C/Currículo 2008

Universidade de Aveiro

Testes de Homogeneidade com Aplicações a Dados Genómicos

Maria Elisabete Fernandes Simões

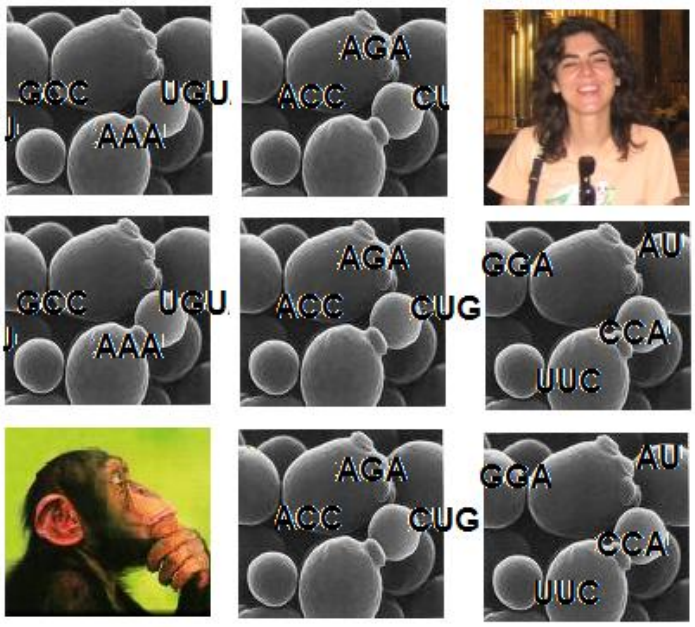
Dissertação de Mestrado em Matemática e Aplicações (2º Ciclo), apresentada ao Departamento de Matemática da Universidade de Aveiro, sob orientação de Prof.ª Doutora Adelaide de Fátima Baptista Valente Freitas.

Universidade de Aveiro Departamento de Matemática

2008

Maria Elisabete Fernandes Simões

Testes de Homogeneidade com Aplicações a Dados Genómicos



Universidade de Aveiro Departamento de Matemática
2008

Maria Elisabete Fernandes Simões

Testes de Homogeneidade com Aplicações a Dados Genómicos

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Prof.ª Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

o júri

presidente Prof.ª Doutora Nélia Maria Marques da Silva

Professora Auxiliar da Universidade de Aveiro.

Prof.ª Doutora Maria do Rosário de Oliveira Silva

Professora Auxiliar do Instituto Superior Técnico da Universidade Técnica de Lisboa.

Prof.ª Doutora Adelaide de Fátima Baptista Valente Freitas

Professora Auxiliar da Universidade de Aveiro (Orientadora).

agradecimentos

À Professora Doutora Adelaide Freitas, pela sua orientação científica, apoio e disponibilidade.

À minha família, em particular, aos meus pais, irmão e avó, pelo enorme apoio e carinho. A eles dedico este trabalho.

Aos imponentes castelos sobre os quais por vezes envaideço.

Às singelas pedras nas quais outras tantas vezes tropeço.

Aos conhecidos, aos vizinhos, aos colegas.

Aos amigos.

palavras-chave Tabela de Contingência; Populações Multinomiais; Testes de Homogeneidade; Estatística de Teste; Nucleótido; Codão; ORFeome.

resumo Na presente dissertação abordamos várias estatísticas de teste para averiguar a existência de homogeneidade entre populações, quando em presença de amostras independentes e amostras emparelhadas. São aqui consideradas, para amostras independentes, a estatística de qui-quadrado de Pearson, a estatística da razão de verosimilhança, uma estatística baseada na medida ϕ-divergência e uma estatística baseada no termo máximo. Para amostras emparelhadas, referimos a estatística de McNemar e a estatística de Stuart-Maxwell.

Em termos de aplicação, o principal objectivo deste trabalho consiste em testar a homogeneidade, no contexto dos codões, de populações consideradas biologicamente próximas, nomeadamente, *Homo sapiens* versus *Pan troglodytes* e *Saccharomyces cerevisiae* versus *Saccharomyces paradoxus*. Comparamos os resultados obtidos usando diferentes estatísticas.

keywords Contingency Table; Multinomial Populations; Homogeneity Tests; Statistic Test; Nucleotide; Codon; ORFeome.

abstract In this dissertation, we analyze different statistics of test in order to test the existence of homogeneity between populations, when in presence of independent samples and matched samples. For independent samples, we consider the Pearson chisquared statistic, the likelihood ratio statistic, a statistic based on ϕ- divergence measure and a statistic based on a maximum term. For matched samples, the McNemar statistic and the Stuart-Maxwell statistic are considered. In terms of applications, the main aim of this work consists in testing the homogeneity, on codon context, of populations considered biologically similar, like Homo sapiens versus Pan Troglodytes and Saccharomyces cerevisiae versus Saccharomyces paradoxus. We compare the results obtained using different statistics.

Conteúdo

Capítulo 1

Introdução

1.1 Contextualização e Conceitos Biológicos

A Matemática, no geral, e a Estatística, em particular, têm acompanhado a Genética desde os seus primórdios, que remontam aos finais do século XIX, com a dedução, por parte de Mendel, dos princípios básicos de transmissão de caracteres hereditários de geração em geração. Os trabalhos de Mendel tiveram como suporte a repetição de experiências com a ervilha-de-cheiro, tendo a interpretação (biológica e matemática) dos resultados sido publicada em 1866.

Mais tarde, entre 1909 e a década de 40, Morgan consegue desvendar o mistério da localização, estabelecendo que os genes são responsáveis pela hereditariedade e se localizam nos cromossomas. Juntamente com os seus colaboradores, usando a frequência estatística de determinados acontecimentos, deduz algumas distâncias entre genes mesmo não conhecendo a sua composição e natureza, dando inicio ao chamado sequenciamento genético que tenta mostrar a disposição dos genes ao longo dos cromossomas.

Em 1944, Avery associa o ácido desoxirribonucleico - ADN - à hereditariedade e desvenda a sua natureza. O gene passa a ser uma entidade física, correspondendo a um fragmento de ADN.

Uma década depois, Waston e Crick, em 1953, estabelecem a estrutura da molécula de ADN, sendo esta formada por duas cadeias enroladas em hélice, constituídas por

sequências de nucleótidos: adenina (A), citosina (C), guanina (G) e uracilo (U)[1] como ilustra a Figura 1.1[2]. Por simplificação, o nucleótido é identificado pela sua base química nitrogenada.

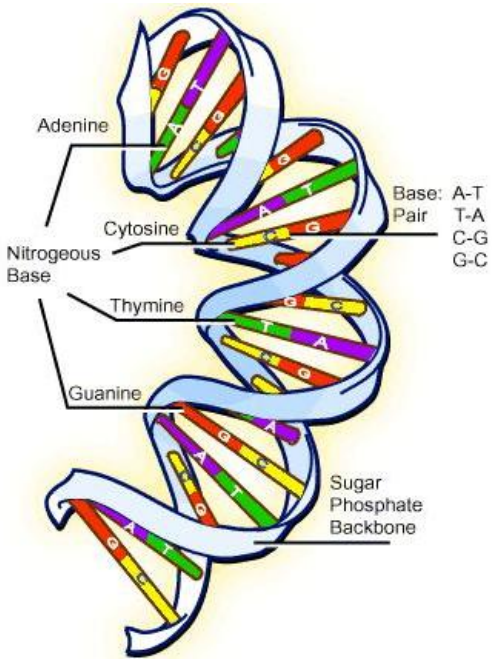


Figura 1.1: Estrutura do ADN - Para além de bases nitrogenadas, a molécula de ADN é ainda formada por ácido fosfórico e pela desoxirribose.

A descoberta do ADN, nos meados do século XX, veio demonstrar-nos que afinal a enorme complexidade e variabilidade de seres vivos se deve à simples combinação de quatro moléculas.

Vários esforços têm sido feitos no sentido de descodificar as mensagens hereditárias "escritas"com o alfabeto de quatro letras: A, C, G e U.

O sequenciamento do ADN de inúmeras espécies, sobretudo bactérias e leveduras, mas também da mosca da fruta, do arroz, e mais recentemente, do homem, entre outros, tem gerado um gigantesco conjunto de dados. O código genético do homem, por exemplo, é composto por cerca de seis mil milhões de letras.

Após o sequenciamento do genoma de um número considerável de seres vivos, ao qual

[1]Na realidade, no ADN encontramos o nucleótido Timina (T) que, no ARN-mensageiro, corresponde ao nucleótido Uracilo.

[2]Retirada da página da Internet com o seguinte endereço: legion.geleia.net/AP/tema1.html.

não se pode alhear o grande desenvolvimento informático registado nas últimas décadas, surge a necessidade de procedermos ao tratamento dessa enorme quantidade de dados aparentemente sem qualquer regularidade.

O tratamento das sequências genéticas tem sido realizado recorrendo a algoritmos que permitam organizar a informação e também, por outro lado, entre outras áreas, recorrendo à teoria da probabilidade e da estatística na tomada de decisões que assentam em modelos probabilísticos, de forma a extrair informação dessas sequências. Para tal tratamento estatístico torna-se fundamental a interacção entre Biólogos, Informáticos e Matemáticos.

Hoje em dia sabemos que as sequências produzidas por combinações de nucleótidos, as quais formam num todo o também denominado texto genético, são compostas por duas partes: uma zona codificada e uma zona não codificada. A zona codificada diz respeito a subsequências ligadas à produção de proteínas. Desconhece-se todavia a funcionalidade da zona não codificada.

Cada sequência do texto genético, com a sua zona codificada associada a uma dada proteína, encontra-se contida num gene. Um gene contém parte codificada e não codificada da sequência de ADN. O conjunto de todos os genes de uma espécie constítui o genoma dessa espécie.

As sequências codificadas do texto genético (que constituem no seu todo o denominado ORFeoma - Open Reading Frame do genoma) são definidas por sequências de nucleótidos agrupadas três a três consecutivamente. Cada tripleto de nucleótidos define um codão, cada codão codifica um aminoácido, a unidade básica na construção da proteína. Cada sequência codificada é iniciada com o codão AUG (denominado codão de iniciação) e termina com um dos codões UAA, UAG ou UGA (denominados codões de terminação). Assim, existem $4^3 = 64$ codões distintos, sendo $61 \times 64 = 3904$ o número de possíveis pares de codões justapostos distintos.

Numa sequência codificada, dado um codão fixo, diz-se que o codão que o antecede está na posição 5' e o codão que o sucede na posição 3', sendo a leitura destas sequências, para a constituição das proteínas, feita da posição 5' para a posição 3'.

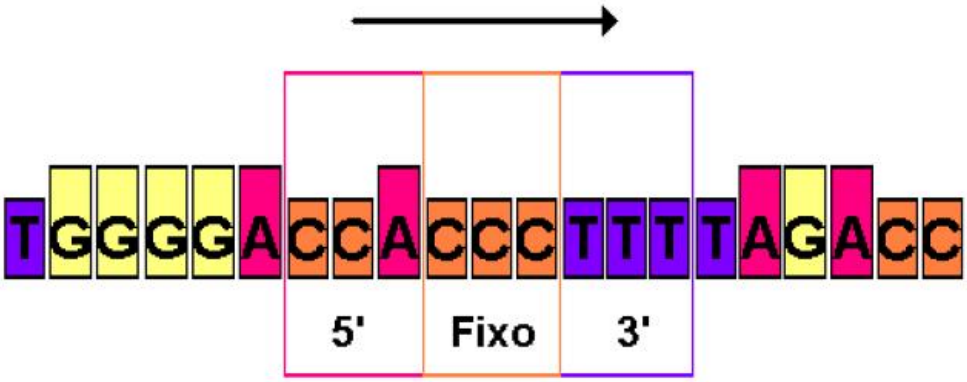


Figura 1.2: Esquema do sentido da leitura da sequência codificada para a constituição das proteínas.

Tendo por objectivo desvendar propriedades inerentes ao desenvolvimento dos seres vivos, acredita-se que o estudo de propriedades intrísecas aos sequenciamentos terá implicações tão fundamentais como a compreensão dos mecanismos da vida, os parentescos entre espécies, o tratamento de doenças genéticas, entre muitos outros.
Um dos temas de interesse na investigação, prende-se com a identificação de regras que governam o sequenciamento dos codões no ORFeoma. Nesse propósito inserem-se os projectos POCTI/BME/39030/2001 (da FCT): "Developing new tools for genome analysis", da Human Frontier Science Program: "mRNA mistranslation in yeast"e PTDC/MAT/72974/2006 (da FCT): "New statistical methodologies for analysis DNA microarrays data", que envolvem investigadores da Universidade de Aveiro, estando a presente dissertação enquadrada nos trabalhos em desenvolvimento naquele último projecto.

1.2 Objectivos Gerais e Organização da Dissertação

A comparação do texto genético de várias espécies revela-se na realidade muito importante, com vista, por exemplo, a compreender ou reconstituir, pelo menos parcialmente, a história da evolução dos seres vivos, a avaliar com precisão o parentesco de um grupo de espécies, a investigar propriedades genéticas comuns ou discordantes entre espécies, etc.

É neste âmbito que surge a motivação para a abordagem de testes estatísticos de homogeneidade na comparação de contextos genéticos.

Esta dissertação, que se insere no projecto PTDC/MAT/72974/2006 (da FCT), tem como principal objectivo contribuir para a investigação sobre a existência de homogeneidade entre pares de espécies tidas biologicamente como semelhantes, o *Homo sapiens* e o *Pan troglodytes* e a *Saccharomyces cerevisiae* e a *Saccharomyces paradoxus*, ao nível do contexto dos codões e nucleótidos. Propomo-nos ainda, pesquisar sobre os codões ou nucleótidos que se apresentem como fortes candidatos à falta de homogeneidade que eventualmente os testes estatísticos possam levar a concluir.

Em [12], no âmbito do projecto POCTI/BME/39030/2001, foi desenvolvido um programa informático chamado *Anaconda*, que permite construir tabelas de contingência obtidas por contagem de pares de símbolos genéticos justapostos, quer sejam eles codões, nucleótidos ou aminoácidos, a partir da leitura de sequências de ORFeomas extraídas de bases de dados públicas. A enorme quantidade de informação contida nas sequências codificadas pode ser organizada em tabelas de contingência de pares de símbolos genómicos para os dois tipos de leituras, a 3' e a 5'.

No nosso estudo, as sequências genómicas das espécies de interesse foram extraídas na base de dados pública situada em www.ensembl.org e processadas no *Anaconda*. Para análise consideramos as tabelas de contingência relativas à leitura 3' por nos parecerem de leitura mais natural. Algumas destas tabelas encontram-se reproduzidas em anexo.

Para além desta introdução, a dissertação é composta por mais quatro capítulos organizados do seguinte modo.

No segundo capítulo, começamos por descrever o modelo probabilístico associado às tabelas de contingência em testes de homogeneidade para amostras independentes (de populações multinomiais), bem como as estimativas de máxima verosimilhança dos parâmetros. Seguidamente, abordaremos quatro estatísticas de teste, possíveis de serem aplicadas a amostras de grandes dimensões, para testar a existência de homogeneidade em populações multinomiais, nomeadamente, a Estatística de Pearson, a Estatística da Razão de Verosimilhança, uma estatística baseada na medida de ϕ-divergência e uma outra baseada no termo máximo das componentes de uma partição

da Estatística de Pearson. Estas estatísticas de teste são aplicadas a amostras de grandes dimensões, relativas a dados genómicos, no quarto capítulo.

No terceiro capítulo, direccionamos para a situação de termos amostras emparelhadas. Nessas condições, analisamos a estatística de McNemar usada para testar o problema da homogeneidade marginal em tabelas 2×2 e a estatística de Stuart-Maxwell em tabelas de contingência $r \times r$.

No quarto capítulo, são aplicadas as estatísticas de testes estudadas no Capítulo 2, no sentido de aferir sobre a existência de homogeneidade entre dois pares de espécies tidas como semelhantes: *Homo sapiens* versus *Pan troglodytes* e *Saccharomyces cerevisiae* versus *Saccharomyces paradoxus*. Estudamos a homogeneidade em termos de distribuição de alguns contextos genómicos e em termos de preferência e preterência de pares de codões consecutivos nos ORFeomas.

Por último, reservamos o quinto capítulo para sumariar os resultados obtidos e apontar algumas observações resultantes do nosso estudo.

Capítulo 2

Testes de Homogeneidade em Amostras Independentes

2.1 Modelo Probabilístico

Os indivíduos de uma dada população podem ser classificados em categorias ou classes, segundo determinado critério. Tal classificação consiste em detectar a categoria a que cada indivíduo pertence, devendo as categorias serem exaustivas e mutuamente exclusivas, isto é, qualquer indivíduo pertencer a uma e uma só categoria.

Para estudar dados categorizados procedemos ao estudo das frequências absolutas ou relativas de cada categoria. Assim, perante uma amostra de dados categorizados, efectuamos a contagem do número de observações em cada categoria, ou seja, calculamos as suas frequências observadas, organizadas, usualmente, em tabelas de contingência.

Considerando A e B duas características (variáveis nominais) de uma determinada população, subdivididas em r e c categorias designadas por $A_1,\ldots,A_r$ e $B_1,\ldots,B_c$, respectivamente, a tabela de contingência que resulta da classificação de n observações ou indivíduos nas $r \times c$ categorias cruzadas tem a forma da Tabela 2.1.

De acordo com a notação introduzida na Tabela 2.1, note-se que:

$$n_{i.} = \sum_{j=1}^{c} n_{ij}\ , \quad i = 1, ..., r\ , \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}\ , \quad j = 1, ..., c\ ,$$

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} \ ,$$

onde n, a dimensão da amostra, se supõe fixa.

	B_1	B_2	$\ldots$	B_c	TotalMarginal
A_1	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	$n_{r.}$
Total Marginal	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.c}$	n

Tabela 2.1: Forma Geral de uma Tabela de Contingência $r \times c$ - A expressão n_{ij} representa o número de observações pertencentes à categoria A_i de A e à categoria B_j de B, $n_{i.}$ o total de observações na categoria A_i da variável A e $n_{.j}$ o total de observações na categoria B_j da variável B, estes últimos designados de totais marginais.

Neste caso, em que se parte de uma amostra em que o número total de indivíduos n está fixo, o que se pretende estudar é a independência entre as variáveis A e B.

Se considerarmos apenas uma característica A subdividida em r categorias, designadas por $A_1,\ldots,A_r$, sobre c populações distintas, $B_1,\ldots,B_c$, a tabela que resulta da classificação de $n_{.j}$ indivíduos da população B_j, com $j = 1,\ldots c$, pelas r categorias da variável A é também designada de tabela de contingência $r \times c$ e tem a forma da Tabela 2.1, sendo os totais das colunas valores prefixados.

Neste segundo caso, o que se pretende averiguar é a existência de homogeneidade entre as c populações, com base em c amostras independentes de tamanho $n_{.1}, \ldots, n_{.c}$ extraídas, cada uma, de uma das c populações (respectiva).

Será sobre o problema da homogeneidade de populações que se irá centrar o nosso estudo.

No entanto, o procedimento de testar a homogeneidade é equivalente a testar a independência em tabelas de contingência $r \times c$. A diferença encontra-se no esquema

de amostragem dos dados e, consequentemente, no modelo probabilístico associado à tabela.

Consideremos então c amostras independentes de c populações, $B_1,\ldots,B_c$, cujos seus elementos se distribuem por r classes ou categorias, $A_1,\ldots,A_r$, de uma variável A (variável linha), em que cada população segue uma distribuição multinomial, sendo $(p_{11},\ldots,p_{r1})$, $(p_{12},\ldots,p_{r2}),\ldots,(p_{1c},\ldots,p_{rc})$ os r parâmetros desconhecidos das populações $B_1,\ldots,B_c$, respectivamente, onde p_{ij} é a probabilidade de um elemento da população B_j pertencer à categoria A_i, $i=1,...,r$, e $\sum_{i=1}^{r} p_{ij} = 1$, para $j = 1,...,c$. Note-se que, para cada j, conhecidos $r-1$ parâmetros, a fórmula $\sum_{i=1}^{r} p_{ij} = 1$ permite calcular o parâmetro restante.

Testar a homogeneidade destas c populações, relativamente à classificação nas r categorias da variável A, equivale a verificar se as proporções de elementos pertencentes a cada uma das categorias da variável A são idênticas para cada uma das populações. Portanto, estamos perante o problema genérico de testar a hipótese nula

$$H_0 : \; p_{i1} = p_{i2} = \ldots = p_{ic} = p_i^{\circ} \quad , \quad i=1,\ldots,r, \tag{2.1}$$

contra a hipótese alternativa

$$H_1 : \exists_{i,j,k} : \quad p_{ij} \neq p_{ik} \; . \tag{2.2}$$

Seja N_{ij} a variável aleatória que representa o número de observações pertencentes à categoria A_i retiradas da população B_j, para $i = 1,...,r$ e $j = 1,...,c$, e $N_{i.} = N_{i1} + N_{i2} + \ldots + N_{ic}$.

Recolhidas as c amostras, n_{ij} e $n_{i.}$ representam as concretizações das variáveis N_{ij} e $N_{i.}$, respectivamente, correspondendo às frequências observadas das amostras com dimensões dimensões $n_{.1}$,$n_{.2}$,..., $n_{.c}$ obtidas das populações $B_1,\ldots,B_c$, respectivamente, e $n = n_{.1} + n_{.2} + \ldots + n_{.c}$.

Nestas condições, para cada população B_j, $j = 1,\ldots,c$, o vector aleatório $(N_{1j}, N_{2j},\ldots,N_{rj})^t$ tem distribuição multinomial, de parâmetros $(n_{.j},\ \mathbf{p}_j)$, onde $\mathbf{p}_j = (p_{1j},\ldots,$

$p_{rj})^t$. Consequentemente,

$$P(N_{1j} = n_{1j}, N_{2j} = n_{2j}, \ldots, N_{rj} = n_{rj}) = \frac{n_{.j}!}{\prod_{i=1}^{r} n_{ij}!} \prod_{i=1}^{r} p_{ij}^{n_{ij}} \ ,$$

com $n_{ij} = 0, \ldots, n_{.j}$ e $\sum_{i=1}^{r} n_{ij} = n_{.j}$.

2.2 Estimativas de Máxima Verosimilhança

Uma estimativa pontual do parâmetro $\mathbf{p}_j$ pode ser obtido pelo método da máxima verosimilhança.

Teorema 2.2.1 *Se o vector aleatório $(N_{1j}, N_{2j}, \ldots, N_{rj})^t$ segue uma distribuição multinomial de parâmetros $(n_{.j}, \mathbf{p}_j)$, as estimativas de máxima verosimilhança de p_{ij} são dadas por*

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_{.j}} \quad , \quad i = 1, \ldots, r \quad , \quad j = 1, \ldots, c.$$

Se for válida a hipótese (2.1), as estimativas de máxima verosimilhança de $p_{ij} = p_i^{\circ}$ serão dada por

$$\widehat{p}_{ij} = \widehat{p_i^{\circ}} = \frac{n_{i.}}{n} \quad , \quad i = 1, \ldots, r \ .$$

Prova 2.2.1 *Considerando $(n_{1j}, n_{2j}, \ldots, n_{rj})^t$ uma realização do vector aleatório $(N_{1j}, N_{2j}, \ldots, N_{rj})^t$, a sua função de verosimilhança é dada por*

$$L(p_{1j}, \ldots, p_{rj}; n_{1j}, \ldots, n_{rj}) = P(N_{1j} = n_{1j}, N_{2j} = n_{2j}, \ldots, N_{rj} = n_{rj}) = \frac{n_{.j}!}{\prod_{i=1}^{r} n_{ij}!} \prod_{i=1}^{r} p_{ij}^{n_{ij}}.$$

A estimativa de máxima verosimilhança de $\mathbf{p}_j$ é dada pelo vector $(\widehat{p}_{1j}, \ldots, \widehat{p}_{rj})^t$ que maximiza a função $L(p_{1j}, \ldots, p_{rj}; n_{1j}, \ldots, n_{rj})$.

Aplicando logaritmos na função de verosimilhança obtemos:

$$\ln L(p_{1j}, \ldots, p_{rj}; n_{1j}, \ldots, n_{rj}) = \ln \ n_{.j}! - \sum_{i=1}^{r} \ln \ n_{ij}! + \sum_{i=1}^{r} n_{ij} \ \ln \ p_{ij}.$$

A primeira e a segunda parcelas do segundo membro são constantes, pois não dependem dos p_{ij}. Dado pretender-se estudar os maximizantes da função de verosimilhança, recorrer-se-á ao

cálculo dos zeros das derivadas para detectar a existência de máximos, tendo apenas interesse o estudo da terceira parcela.

Sendo L *uma função de* $(p_{1j}, \ldots, p_{r-1j})$, *pois* $p_{rj} = 1 - (p_{1j} + \ldots + p_{r-1j})$, *tem-se que* $\dfrac{\partial p_{rj}}{\partial p_{ij}} = -1$, $i = 1, \ldots, r-1$. *Assim,*

$$\frac{\partial \ln p_{rj}}{\partial p_{ij}} = \frac{1}{p_{rj}} \frac{\partial p_{rj}}{\partial p_{ij}} = -\frac{1}{p_{rj}}.$$

Derivando $\ln L$ *e igualando a zero, para* $i = 1, 2, \ldots, r-1$, *vem:*

$$\begin{aligned} \frac{\partial \ln L}{\partial p_{ij}} = 0 \Leftrightarrow \frac{\partial (n_{ij} \ln p_{ij})}{\partial p_{ij}} + \frac{\partial (n_{rj} \ln p_{rj})}{\partial p_{ij}} = 0 \\ \Leftrightarrow \frac{n_{ij}}{p_{ij}} - \frac{n_{rj}}{p_{rj}} = 0 \\ \Leftrightarrow \frac{n_{ij}}{p_{ij}} = \frac{n_{rj}}{p_{rj}}, \end{aligned}$$

donde as soluções de máxima verosimilhança satisfazem $\dfrac{\widehat{p}_{ij}}{\widehat{p}_{rj}} = \dfrac{n_{ij}}{n_{rj}}$. *Consequentemente,*

$$\sum_{i=1}^{r} \widehat{p}_{ij} = \frac{\widehat{p}_{rj} \sum_{i=1}^{r} n_{ij}}{n_{rj}} = \frac{\widehat{p}_{rj} n_{.j}}{n_{rj}}.$$

Mas, por outro lado, $\sum_{i=1}^{r} \widehat{p}_{ij} = 1$, *donde resulta que*

$$\widehat{p}_{rj} = \frac{n_{rj}}{n_{.j}} \ ,$$

e assim,

$$\widehat{p}_{ij} = \frac{n_{rj}}{n_{.j}} \frac{n_{ij}}{n_{rj}} = \frac{n_{ij}}{n_{.j}}.$$

Se for válida a hipótese de homogeneidade (2.1),

$$\widehat{p}_{ij} = \widehat{p_i^{\circ}} = \frac{\sum_{j=1}^{c} n_{ij}}{n} = \frac{n_{i.}}{n} \quad , \quad i = 1, \ldots, r \ .$$

□

Sendo p_{ij} a probabilidade de um elemento da população B_j pertencer à categoria A_i, recolhida uma amostra de $n_{.j}$ elementos da população B_j, espera-se que entre esses n_{ij} elementos existam $n_{.j}p_{ij}$ elementos na categoria A_i.

Sendo as estimativas de máxima verosimilhança de p_{ij} dadas por $\widehat{p}_{ij} = \dfrac{n_{ij}}{n_{.j}}$, as estimativas de máxima verosimilhança das frequências esperadas serão dadas por:

$$\widehat{e}_{ij} = n_{.j}\widehat{p}_{ij} = n_{.j}\frac{n_{ij}}{n_{.j}} = n_{ij} \quad , \qquad i = 1, \ldots, r \quad , \quad j = 1, \ldots, c.$$

Sob a hipótese H_0,

$$\widehat{e}_{ij} = n_{.j}\widehat{p}_{ij} = \frac{n_{.j}n_{i.}}{n} \quad , \quad i = 1, \ldots, r \quad , \quad j = 1, \ldots, c.$$

2.3 Estatística de Pearson

Um dos métodos mais usados para testar a hipótese de homogeneidade em distribuições multinomiais é tomar, como estatística de teste, a chamada estatística do qui-quadrado de Pearson, que é dada por:

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(N_{ij-}\,\widehat{e}_{ij}\right)^2}{\widehat{e}_{ij}}\Bigg|_{H_0} = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(N_{ij} - \dfrac{n_{.j}N_{i.}}{n}\right)^2}{\dfrac{n_{.j}N_{i.}}{n}}, \tag{2.3}$$

que segue assimptoticamente uma distribuição de qui-quadrado com $(r-1)(c-1)$ graus de liberdade [2,4].

Quando não ocorre homogeneidade, é natural que as frequências observadas n_{ij} sejam substancialmente diferentes das frequências e_{ij} que esperamos observar quando existe homogeneidade. Então, devemos rejeitar a hipótese de homogeneidade quanto o valor observado para a estatística de teste, sob H_0, tem um valor relativamente grande, isto é, quando $\chi^2_{obs} > k$. Dado um nível de significância α, o valor crítico k é determinado através de α =P($\chi^2 > k \mid H_0$ é verdadeira). Assim, concluí-se que k é o quantil de ordem $1-\alpha$ de uma distribuição de qui-quadrado com $(r-1)(c-1)$ graus de liberdade, que denotamos por $\chi^2_{(r-1)(c-1);1-\alpha}$.

Em resumo, devemos rejeitar a hipótese H_0 de homogeneidade das c populações, ao nível de significância α, se

$$\chi^2_{obs} > \chi^2_{(r-1)(c-1);1-\alpha},$$

onde χ^2_{obs} indica o valor observado para a estatística de Pearson, obtido com base na amostra considerada.

Uma decisão em termos de p-value será a de rejeitar a hipótese H_0 de homogeneidade das c populações, ao nível de significância α, se

$$p\text{-value} = P(\chi^2 > \chi^2_{obs} | H_0 \text{ é verdadeira}) < \alpha.$$

No caso particular de $r = 2$ e $c = 2$, estamos perante uma tabela de contingência 2×2 com a forma da Tabela 2.2.

	B_1	B_2	Total Marginal
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
Total Marginal	$n_{.1}$	$n_{.2}$	n

Tabela 2.2: Forma Geral de uma Tabela de Contingência 2×2.

Neste caso particular é fácil verificar, através de cálculos algébricos simples, que (2.3) se reduz à expressão:

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{11} + n_{12})(n_{11} + n_{21})(n_{12} + n_{22})(n_{21} + n_{22})}\,. \tag{2.4}$$

A distribuição de frequências de qualquer tabela de contingência é uma distribuição discreta. Ao realizarmos um teste com a estatística do qui-quadrado, estamos a aproximar a distribuição de frequências a uma distribuição contínua. A fim de melhorar a aproximação, Yates (1934) sugeriu o uso de uma correcção de continuidade, que logo tomou o seu nome, que consiste em adicionar 0.5 ao desvio $(n_{ij} - e_{ij})$ quando esta diferença é negativa, e a subtrair-lhe 0.5 no caso de ser positiva, antes dos seus quadrados serem utilizados no cálculo do valor da estatística.

No caso de uma tabela 2×2, o valor da estatística de Pearson (2.4), com a correcção de Yates, é calculada através da seguinte expressão:

$$\chi^2 = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - n/2)^2}{(n_{11} + n_{12})(n_{11} + n_{21})(n_{12} + n_{22})(n_{21} + n_{22})}.$$

2.3.1 Contagem dos Graus de Liberdade

O cálculo da estatística de Pearson, χ^2, requer o conhecimento de $r \times c$ parâmetros desconhecidos, correspondentes às frequências esperadas $e_{ij} = n_{.j}p_{ij}$ as quais estimadas sob a validade de H_0.

Uma vez que as frequências esperadas dos totais marginais $e_{i.}$ e $e_{.j}$ coincidem com as frequências observadas $n_{i.}$ e $n_{.j}$, respectivamente, estes são fixados e alguns daqueles $r \times c$ parâmetros desconhecidos não serão livres na sua determinação.

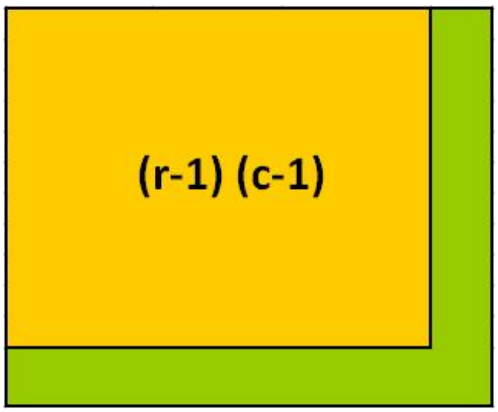


Figura 2.1: Das $r \times c$ frequências esperadas, apenas $(r-1)(c-1)$ são livres.

Na Tabela 2.1, sendo conhecida a dimensão de cada amostra, correspondente ao total de cada coluna, a distribuição multinomial associada às frequências referentes às r categorias da variável A é feita com uma restrição, pois conhecidos $r-1$ parâmetros, a fórmula $\sum_{i=1}^{r} p_{ij} = 1$ permite calcular o último parâmetro. Assim, em cada uma das c colunas, temos livres apenas $r-1$ frequências.

Mas, conhecendo os r totais das linhas, temos então como independentes as frequências de $c-1$ amostras, cada uma com $r-1$ frequências independentes. Donde podemos concluir, que o total de graus de liberdade será $(r-1)(c-1)$.

2.3.2 Restrições na Aplicação da Estatística do Qui-Quadrado

Vários autores sugerem que o valor mínimo de e_{ij} deve ser superior a dez, numa tabela de contingência 2×2, e superior a cinco se o número de graus de liberdade for não inferior a dois [2,4]. Quando e_{ij} desce abaixo daqueles valores habituais de referência, é usual propor a junção de classes em que tal acontece com outras, a fim de se obter novas classes com maiores frequências esperadas. Porém, este procedimento nem sempre é aconselhável visto afectar a aleatoriedade da amostra, podendo eventualmente haver perda de informação na fusão de categorias [4].

Uma outra limitação conhecida na aplicação da estatística de Pearson é o facto de esta requerer amostras de grande dimensão, dado a referência à distribuição de qui-quadrado ser assimptótica.

2.3.3 Partições da Estatística de Pearson

Para além do estudo da hipótese (2.1) sobre uma tabela de contingência $r \times c$, é ainda possível fazer uma análise mais pormenorizada da tabela combinando informação de diversas tabelas extraídas da tabela original. Uma das maneiras mais simples, consiste em estudar a importância dos desvios entre as frequências observadas e as esperadas de cada categoria, através de contribuições individuais da estatística.

O saber quais as categorias que mais contribuem para a obtenção de eventuais valores significativos de χ^2 pode resolver-se particionando a estatística. A dificuldade residirá na interpretação dos resultados, como se verá em termos práticos no Capítulo 4 aquando do estudo da homogeneidade num problema concreto.

Lancaster e Irwin (1949) demonstraram que se pode particionar a estatística de Pearson em tantas componentes independentes quanto o número de graus de liberdade da estatística.

Cada componente corresponde uma tabela 2×2, construída a partir da tabela inicial, à qual está associada uma distribuição de qui-quadrado com um grau de liberdade. Os qui-quadrados obtidos correspondem a componentes independentes.

Essas tabelas 2×2 são construídas da forma como está esquematizado na Figura 2.2.

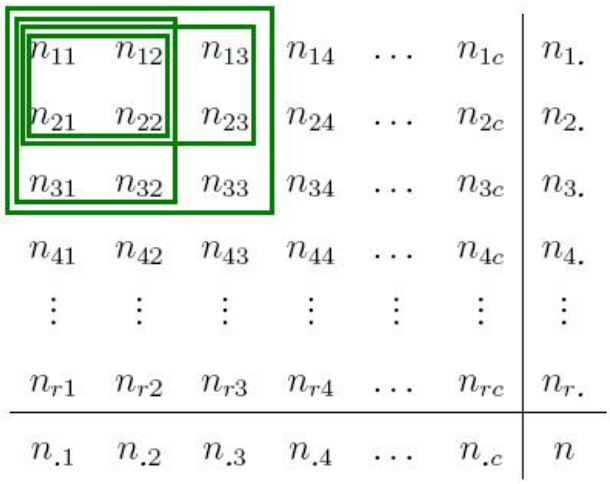


Figura 2.2: Esquema que mostra as quatro tabelas 2×2 obtidas na partição de uma tabela de contingência 3×3.

Por exemplo, numa tabela 3×4 as seis tabelas 2×2 seriam:

n_{11}	n_{12}
n_{21}	n_{22}

$n_{11} + n_{12}$	n_{13}
$n_{21} + n_{22}$	n_{23}

$n_{11} + n_{12} + n_{13}$	n_{14}
$n_{21} + n_{22} + n_{23}$	n_{24}

n_{11}	n_{12}
$+$	$+$
n_{21}	n_{22}
n_{31}	n_{32}

$n_{11} + n_{12}$	n_{13}
$+$	$+$
$n_{21} + n_{22}$	n_{23}
$n_{31} + n_{32}$	n_{33}

$n_{11} + n_{12} + n_{13}$	n_{14}
$+$	$+$
$n_{21} + n_{22} + n_{23}$	n_{24}
$n_{31} + n_{32} + n_{33}$	n_{34}

Na prática, a construção das tabelas 2×2 que definem uma partição da estatística de Pearson dependem da ordem pela qual as categorias estão organizadas na tabela.

Para o caso de uma tabela de contingência $r \times c$, Kimball (1954) desenvolveu fórmulas para a partição da estatística de Pearson através de $(r-1)(c-1)$ estatísticas, cada uma com distribuição de qui-quadrado com um grau de liberdade.

Por exemplo, as componentes independentes associadas às quatro tabelas 2×2 resultantes da partição de uma tabela de contingência 3×3, são dadas pelas seguintes fórmulas:

$$
\begin{aligned}
\chi_1^2 &= \frac{n[n_{2.}(n_{.2}n_{11} - n_{.1}n_{12}) - n_{1.}(n_{.2}n_{21} - n_{.1}n_{22})]^2}{n_{1.}n_{2.}n_{.1}n_{.2}(n_{1.} + n_{2.})(n_{.1} + n_{.2})}, \\
\chi_2^2 &= \frac{n^2[n_{23}(n_{11} + n_{12}) - n_{13}(n_{21} + n_{22})]^2}{n_{1.}n_{2.}n_{.3}(n_{1.} + n_{2.})(n_{.1} + n_{.2})}, \\
\chi_3^2 &= \frac{n^2[n_{32}(n_{11} + n_{21}) - n_{31}(n_{12} + n_{22})]^2}{n_{3.}n_{.1}n_{.2}(n_{1.} + n_{2.})(n_{.1} + n_{.2})}, \\
\chi_4^2 &= \frac{n[n_{33}(n_{11} + n_{12} + n_{21} + n_{22}) - (n_{13} + n_{23})(n_{31} + n_{32})]^2}{n_{3.}n_{.3}(n_{1.} + n_{2.})(n_{.1} + n_{.2})}.
\end{aligned}
\tag{2.5}
$$

As fórmulas simplificadas de Kimball sobre tabelas de contingência $r \times 2$, são dadas por:

$$
\chi_t^2 = \frac{n^2 \left(N_{t+1,2} \sum_{i=1}^{t} N_{i1} - N_{t+1,1} \sum_{i=1}^{t} N_{i2} \right)^2}{n_{1.}n_{2.}N_{t+1,.} \sum_{i=1}^{t} N_{i.} \sum_{i=1}^{t+1} N_{i.}}, \quad t = 1, \ldots, r-1. \tag{2.6}
$$

2.4 Estatística da Razão de Verosimilhança

Uma alternativa à utilização da estatística do qui-quadrado de Pearson é tomar a chamada estatística da razão de verosimilhança.

Sendo a função de verosimilhança a probabilidade da amostra, vista como uma função dos parâmetros assumindo os valores da amostra conhecidos, a estatística de teste da razão de verosimilhança Λ é definida pelo seguinte quociente:

$$
\Lambda = \frac{\text{máxima verosimilhança quando os parâmetros satisfazem } H_0}{\text{máxima verosimilhança quando os parâmetros não são restritos}}.
$$

De acordo com a demonstração e enunciado do Teorema 2.2.1, o núcleo da função de verosimilhança é

$$\prod_{j=1}^{c}\prod_{i=1}^{r} p_{ij}^{n_{ij}},$$

sob a hipótese de homogeneidade (2.1), os valores dos parâmetros p_{ij} que maximizam a função de verosimilhança são dados pelas estimativas de máxima verosimilhança, $\widehat{p}_{ij} = \dfrac{n_{i.}}{n}$ e, no caso geral, os valores de p_{ij} serão dados pelas estimativas de máxima verosimilhança sob a condição de que H_0 possa ou não ser verdadeira, ou seja, por $\widehat{p}_{ij} = \dfrac{n_{ij}}{n_{.j}}$. Assim, a razão de verosimilhança Λ é igual a

$$\Lambda = \frac{\displaystyle\prod_{j=1}^{c}\prod_{i=1}^{r}\left(\frac{n_{i.}}{n}\right)^{n_{ij}}}{\displaystyle\prod_{j=1}^{c}\prod_{i=1}^{r}\left(\frac{n_{ij}}{n_{.j}}\right)^{n_{ij}}} = \prod_{j=1}^{c}\prod_{i=1}^{r}\left(\frac{\frac{n_{i.}}{n}}{\frac{n_{ij}}{n_{.j}}}\right)^{n_{ij}}.$$

Aplicando logaritmos, vem:

$$\ln\Lambda = \sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\ \ln\frac{\left(\frac{n_{i.}}{n}\right)}{\left(\frac{n_{ij}}{n_{.j}}\right)}.$$

Em geral, a estatística a usar é $-2\ln\Lambda$, sendo denotada por G^2. Assim temos

$$G^2 = 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\ \left(\ln\left(\frac{n_{ij}}{n_{.j}}\right) - \ln\left(\frac{n_{i.}}{n}\right)\right),$$

a qual segue assimptoticamente uma distribuição de qui-quadrado com $(r-1)(c-1)$ graus de liberdade [9].

À semelhança do que acontece com a estatística de Pearson, os valores possíveis da estatística G^2 são não negativos, tomando o valor mínimo zero quando os valores observados n_{ij} são iguais aos valores esperados $e_{ij} = \dfrac{n_{i.}n_{.j}}{n}$.

Quando H_0 é falsa, a razão de verosimilhança Λ tenderá a tomar valores abaixo de um, onde o logaritmo é negativo, e assim $G^2 = -2\ln\Lambda$ tenderá a tomar um valor positivo

relativamente grande.

As estatísticas de Pearson e da razão de verosimihança, embora sendo estatísticas de teste distintas, usualmente levam às mesmas conclusões e partilham algumas propriedades como, por exemplo, ambas requerem amostras de grandes dimensões ou o facto do seu valor não depender da ordenação das linhas e das colunas na tabela de contingência por ambas trabalharem com dados nominais.

2.5 Estatística Baseada na ϕ-Divergência

As estatísticas de Pearson e da razão de verosimilhança são casos particulares de uma família de estatísticas definidas em termos da chamada medida de ϕ-divergência.

A medida de ϕ-divergência, introduzida independentemente por Csiszár (1967) e Ali e Silvey (1966), é definida do seguinte modo:

Definição 2.5.1 *Seja Φ o conjunto de todas as funções convexas $\phi(t)$, $t > 0$, tais que*

- $\phi(0) = 1$;
- $0\phi(\frac{0}{0}) = 0$;
- $0\phi(\frac{q}{0}) = \lim_{u\to+\infty} \frac{\phi(u)}{u}$;
- $\phi'(1) = 0$[1] ;
- $\phi''(1) > 0$.

Dadas duas distribuições de probabilidade discretas $\boldsymbol{q} = (q_{11}, \ldots, q_{rc})^t$ e $\boldsymbol{s} = (s_{11}, \ldots, s_{rc})^t$, para uma função $\phi \in \Phi$, chama-se medida de ϕ-divergência entre aquelas duas distribuições à função

$$D_\phi(\boldsymbol{q}, \boldsymbol{s}) = \sum_{j=1}^{c}\sum_{i=1}^{r} s_{ij}\phi\left(\frac{q_{ij}}{s_{ij}}\right) .$$

[1]Na definição e nos resultados teóricos desenvolvidos por vários autores [9], a condição $\phi'(1) = 0$ pode ser eliminada pela condição menos restritiva de ϕ ser diferenciável para $x = 1$. Nos casos aqui considerados trabalharemos com funções ϕ tais que $\phi'(1) = 0$.

Estando perante o objectivo de testar a hipótese (2.1) de homogeneidade, esta família de estatísticas de teste de ϕ - divergência é constituída por todas as estatísticas com valores obtidos da forma

$$T_{\phi_1,\phi_2} = \frac{2n}{\phi_1''(1)} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\theta}_{\phi_2})), \tag{2.7}$$

onde $\phi_1, \phi_2 \in \Phi$, $\widehat{\mathbf{p}}$ é o estimador de máxima verosimilhança não restrito, isto é, satisfazendo as hipóteses (2.1) e (2.2), de $\mathbf{p} = (p_{11}, \ldots, p_{rc})$, e dado por

$$\widehat{\mathbf{p}} = \left(\frac{N_{ij}}{n}\right) \quad , \; i = 1, \ldots, r, \; , j = 1, \ldots, c,$$

e $\mathbf{p}^*(\widehat{\theta}_{\phi_2})$ é uma função do parâmetro $\mathbf{p}$ dada por

$$\begin{aligned} \mathbf{p}^*(\mathbf{p}) &= \left[\frac{n_{.1}}{n}, \ldots, \frac{n_{.1}}{n}, \ldots, \frac{n_{.c}}{n}, \ldots, \frac{n_{.c}}{n}\right] [p_{11}, \ldots, p_{rc}]^t \\ &= \left[\frac{n_{.1}}{n}, \ldots, \frac{n_{.1}}{n}, \ldots, \frac{n_{.c}}{n}, \ldots, \frac{n_{.c}}{n}\right] \mathbf{p}^t, \end{aligned}$$

em que $\mathbf{p}$ é estimado usando o denominado estimador de potência-divergência mínimo, $\widehat{\theta}_{\phi_2}$, em vez da estimativa de máxima verosimilhança de $\mathbf{p}_j$, $\widehat{\mathbf{p}_j} = \left(\frac{n_{1j}}{n_{.j}}, \ldots, \frac{n_{rj}}{n_{.j}}\right)^t$, sendo $\mathbf{p}_j = (p_{1j}, \ldots, p_{rj})^t$ o vector dos parâmetros da multinomial associada à população B_j.

Usando o estimador de máxima verosimilhança restrito, a medida de ϕ-divergência entre os vectores $\widehat{\mathbf{p}}$ e $\mathbf{p}^*(\mathbf{p})$ é dada por

$$D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\mathbf{p})) = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{.j}}{n} p_{ij} \phi_1 \left(\frac{n_{ij}}{n_{.j} p_{ij}}\right).$$

Em 2003, Menéndez, J.A. Pardo, L. Pardo e Zografos, em [9], introduziram o conceito de estimador de ϕ-divergência mínimo, para o problema da homogeneidade, generalizando o conceito de estimador de máxima verosimilhança restrito.

Definição 2.5.2 *O valor* $\widehat{\theta}_{\phi_2}$ *verificando*

$$D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\theta}_{\phi_2})) = \inf_{\{\mathbf{p}\,:\, p_{ij}-p_{ic}\,=0,\, i=1,\ldots,r-1,\, j=1,\ldots,c-1\}} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\mathbf{p})),$$

com $\phi_1, \phi_2 \in \Phi$*, denomina-se estimativa de* ϕ*-divergência mínima para o parâmetro* $\mathbf{p} = (p_{11}, \ldots, p_{r1}, \ldots, p_{1c} \ldots, p_{rc})$.

Ao estudarem propriedades assimptóticas, bem como a distribuição assimptótica da estatística de teste T_{ϕ_1,ϕ_2}, estimando os parâmetros usando o estimador de ϕ-divergência mínimo sobre determinadas condições, em vez do estimador de máxima verosimilhança, aqueles autores concluíram que todos os estimadores de ϕ-divergência mínimo, sob a hipótese da homogeneidade, têm o mesmo comportamento assimptótico, isto é, são independentes da função ϕ considerada.
Para a obtenção do estimador de ϕ-divergência mínimo, sob a hipótese de homogeneidade, recorreram a uma importante família de medidas de ϕ-divergência, denominadas de medidas de potência-divergência, introduzida por Cressie e Read (1984), e definida da seguinte forma:

$$\phi_{(\lambda)}(x)^2 = \begin{cases} (\lambda(\lambda+1))^{-1}(x^{\lambda+1} - x) & \text{, para} \quad \lambda \neq 0, \lambda \neq -1 \\ \lim_{\lambda \to 0} \phi_{(\lambda)}(x) & \text{, para} \quad \lambda = 0 \\ \lim_{\lambda \to -1} \phi_{(\lambda)}(x) & \text{, para} \quad \lambda = -1 \end{cases} .$$

Teorema 2.5.1 *A estimativa de potência-divergência mínima,* $\widehat{\theta}_{\phi_2(\lambda)}$*, de* $\mathbf{p}$*, sob a hipótese de homogeneidade (2.1), é dada por*

[2]Podendo, por vezes, tomar outra forma em virtude de $\phi_{(\lambda)}(x)$ e $\psi_{(\lambda)}(x) \equiv \phi_{(\lambda)}(x) - (x-1)(\lambda+1)^{-1}$ definirem a mesma medida de divergência.

$$\widehat{\theta}_{i,\ \phi_2(\lambda)} = \frac{\left(\sum_{j=1}^{c} \frac{n_{ij}^{\lambda+1}}{n_{\cdot j}^{\lambda}}\right)^{\frac{1}{(\lambda+1)}}}{\sum_{i=1}^{r}\left(\sum_{j=1}^{c} \frac{n_{ij}^{\lambda+1}}{n_{\cdot j}^{\lambda}}\right)^{\frac{1}{(\lambda+1)}}}, \quad i = 1, \ldots, r.$$

Observa-se que, tomando $\lambda = 0$, obtém-se a estimativa de máxima verosimilhança, sob a hipótese de homogeneidade,

$$\widehat{\theta}_{i,\ \phi_2(0)} = \frac{n_{i\cdot}}{n}, \quad i = 1, \ldots, r,$$

e tomando $\lambda = 1$, obtém-se a denominada estimativa mínima do qui-quadrado para a homogeneidade, a qual minimiza a estatística de Pearson, dada por

$$\widehat{\theta}_{i,\ \phi_2(1)} = \frac{\left(\sum_{j=1}^{c} \frac{n_{ij}^{2}}{n_{\cdot j}}\right)^{\frac{1}{2}}}{\sum_{i=1}^{r}\left(\sum_{j=1}^{c} \frac{n_{ij}^{2}}{n_{\cdot j}}\right)^{\frac{1}{2}}}, \quad i = 1, \ldots, r.$$

Em relação à distribuição assimptótica da estatística de teste (2.7), T_{ϕ_1,ϕ_2}, em que a primeira função (ϕ_1) corresponde à função que é usada na estatística de teste para testar e a segunda função (ϕ_2) é usada na determinação do estimador, tem-se o seguinte resultado:

Teorema 2.5.2 *A estatística de teste*

$$T_{\phi_1,\phi_2} = \frac{2n}{\phi_1''(1)} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\theta}_{\phi_2})),$$

para testar a hipótese (2.1) de homogeneidade, segue assimptóticamente a distribuição de qui-quadrado com $(c-1)(r-1)$ *graus de liberdade.*

Assim, deve rejeitar-se a hipótese de homogeneidade, a nível de significância α, se

$$T_{\phi_1,\phi_2} > \chi^2_{(r-1)(c-1);\,1-\alpha}.$$

Quando $\phi_2(x) = x\ln x - x + 1$, Gupta et al (2007)[7] indica que $\widehat{\theta}_{\phi_2}$ é igual à estimativa de máxima verosimilhança de $\mathbf{p}$, sob a hipótese nula, ou seja,

$$\widehat{\theta}_{\phi_2} = \widehat{\widehat{\mathbf{p}}} = \left(\frac{n_{1.}}{n}, \ldots, \frac{n_{r.}}{n}, \ldots, \frac{n_{1.}}{n}, \ldots, \frac{n_{r.}}{n}\right). \tag{2.8}$$

Proposição 2.5.1 *A estatística de teste*

$$T_{\phi_1,\phi_2} = \frac{2n}{\phi_1''(1)} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\theta}_{\phi_2})),$$

coincide com:

i) *a estatística de Pearson quando tomamos* $\phi_1(x) = \frac{1}{2}(x-1)^2$ *e* $\phi_2(x) = x\ln x - x + 1$*;*

ii) *a estatística da razão de verosimilhança quando tomamos* $\phi_1(x) = \phi_2(x) = x\ln x - x + 1$.

Prova 2.5.1 *Tendo em conta (2.8) para* $\phi_2(x) = x\ln x - x + 1$*, temos:*

i)

$$\begin{aligned}
T_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\boldsymbol{p}})) &= 2n \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{.j}}{n}\, \widehat{\widehat{p}}_{ij}\, \phi_1\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right) \\
&= 2n \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{.j}}{n}\, \widehat{\widehat{p}}_{ij}\, \frac{1}{2}\left(\frac{n_{ij}}{n_{.j}}\widehat{\widehat{p}}_{ij} - 1\right)^2 \\
&= \frac{2n}{2} \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{.j}}{n}\, \widehat{\widehat{p}}_{ij} \left[\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right)^2 - 2\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}} + 1\right] \\
&= n \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{n n_{.j}\widehat{\widehat{p}}_{ij}} - 2\frac{n_{ij}}{n} + \frac{n_{.j}}{n}\widehat{\widehat{p}}_{ij} \\
&= \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{\frac{n_{i.}n_{.j}}{n}} - 2n_{ij} + \frac{n_{i.}n_{.j}}{n} \quad \textit{(1)}
\end{aligned}$$

dado que $\widehat{\widehat{p}}_{ij} = \dfrac{n_{i.}}{n}$.

Por outro lado,

$$\begin{aligned}\chi^2 \equiv \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} &= \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2 - 2\frac{n_{ij}n_{i.}n_{.j}}{n} + \left(\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \\ &= \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{\frac{n_{i.}n_{.j}}{n}} - 2n_{ij} + \frac{n_{i.}n_{.j}}{n} \quad \textbf{(2)}\end{aligned}$$

Logo (1)=(2).

ii)

$$\begin{aligned}T_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}(\widehat{\mathbf{p}}, \mathbf{p}^*(\widehat{\widehat{\boldsymbol{p}}})) &= 2n\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{.j}}{n}\,\widehat{\widehat{p}}_{ij}\;\phi_1\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right) \\ &= 2n\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{.j}}{n}\,\widehat{\widehat{p}}_{ij}\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\ln\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right) - \frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}} + 1\right) \\ &= 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\ln\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right) - n_{ij} + n_{.j}\widehat{\widehat{p}}_{ij} \\ &= 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\ln\left(\frac{n_{ij}}{n_{.j}\widehat{\widehat{p}}_{ij}}\right) - n_{ij} + \frac{n_{.j}n_{i.}}{n} \\ &= 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\left(\ln\left(\frac{n_{ij}}{n_{.j}}\right) + \ln\left(\frac{1}{\widehat{\widehat{p}}_{ij}}\right)\right) \\ &= 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}\left(\ln\left(\frac{n_{ij}}{n_{.j}}\right) - \ln\left(\frac{n_{i.}}{n}\right)\right)\end{aligned}$$

□

Em [9] é ainda apresentado um pequeno estudo onde se conclui que tomando a medida potência-divergência, a estimativa mínima do qui-quadrado, $\widehat{\theta}_{\phi_{2\,(1)}}$, e a denominada estimativa de Cressie-Read, $\widehat{\theta}_{\phi_{2(\frac{2}{3})}}$, oferecem uma alternativa atractiva à estimativa da máxima verosimilhança.

Outra conclusão, da simulação apresentada por aqueles autores, foi a de que o comportamento da estatística T_{ϕ_1,ϕ_2} depende do tamanho da amostra, obtendo as estimativas $\widehat{\theta}_{\phi_{2(\lambda)}}$ melhores resultados quanto maior for o tamanho da amostra.

Comparando a potência da estatística de teste $T_{\phi_{1(\lambda_1)}\;\phi_{2(\lambda_2)}}$, o estudo concluíu que as estatísticas de teste $T_{\phi_{1(\frac{2}{3})},\;\phi_{2(0)}}$ e $T_{\phi_{1(\frac{2}{3})},\;\phi_{2(1)}}$ são boas alternativas à estatística de teste

de Pearson ($T_{\phi_{1(1)},\,\phi_{2(0)}}$) e à estatística de teste da razão de verosimilhança ($T_{\phi_{1(0)},\,\phi_{2(0)}}$) considerando amostras de pequena e média dimensão.

2.6 Estatística Baseada no Termo Máximo

Uma outra estatística, denominada T_r, para o problema da homogeneidade, sobre duas populações multinomiais, B_1 e B_2, foi proposta em Freitas el at (2005).
Genericamente, consideramos duas amostras independentes organizadas numa tabela de contingência $r \times 2$, e o problema de testar

$$H_0 : p_{i1} = p_{i2} \quad , \quad i = 1, \ldots, r \quad ,$$

contra

$$H_1 : \exists i : \quad p_{i1} \neq p_{i2} \quad .$$

A estatística de teste T_r é definida como o máximo, em vez da soma, das $r-1$ componentes mutuamente independentes, χ^2_1, χ^2_2, $\ldots$, χ^2_{r-1}, cada uma assimptoticamente distribuída por um qui-quadrado com um grau de liberdade, da partição da estatística de Pearson formulada por Kimball, que conforme (2.6), são dadas por:

$$\chi^2_i = \frac{n^2(N_{i+1,2}(N_{11} + N_{12} + \ldots + N_{i1}) - N_{i+1,1}(N_{12} + N_{22} + \ldots + N_{i2}))^2}{n_{1.}n_{2.}N_{i+1,.}(N_{1.} + N_{2.} + \ldots + N_{i.})(N_{1.} + N_{2.} + \ldots + N_{i+1,.})}.$$

A aparente vantagem da estatística T_r, reside no facto das $r-1$ componentes poderem identificar as categorias responsáveis por uma eventual rejeição da homogeneidade. Porém, como veremos em casos práticos estudados no quarto capítulo, essa identificação levanta questões.
Observemos que, de acordo com o indicado atrás sobre a fórmula (2.6), a estatística χ^2_i é a estatística de Pearson usada para testar a hipótese de homogeneidade $H_{0,\,i}$: $p_{i+1,1} = p_{i+1,2}$, sob uma tabela de contingência 2×2 obtida da tabela de contingência $r \times 2$ inicial da forma descrita na Figura 2.2 para $c = 2$, nomeadamente: a primeira

linha é obtida somando as primeiras i categorias da tabela de contingência $r \times 2$, e a sua segunda linha é a linha $i+1$ da tabela de contingência $r \times 2$.

A estatística de teste $T_r = \max(\chi_1^2, \chi_2^2, \ldots, \chi_{r-1}^2)$ surge do facto da hipótese H_0 de homogeneidade entre B_1 e B_2 corresponder a uma intersecção de $r-1$ hipóteses nulas, $H_0 = \bigcap_{i=1}^{r-1} H_{0,\,i}$. A hipótese $H_{0,\,i}$ é relativa à homogeneidade de duas sub-populações de B_1 e B_2 com duas categorias: uma comum às populações originais e a outra obtida por combinação de categorias de uma forma apropriada.

Sendo

$$T_r = \max(\chi_1^2, \chi_2^2, \ldots, \chi_{r-1}^2),$$

então

$$\lim_{n \to +\infty} P(T_r \leq t) = F^{r-1}(t) \quad , t \in \mathbb{R}, \tag{2.9}$$

onde F é a função de distribuição de qui-quadrado com um grau de liberdade.

Da Teoria de Valores Extremos, sabemos quais as possíveis distribuições limite para o máximo, convenientemente normalizado, nomeadamente, temos o seguinte resultado:

Teorema 2.6.1 *Se a função distribuição $F^r(a_r x + b_r)$, com constantes de normalização $b_r \in \mathbb{R}$ e $a_r > 0$, converge para uma função de distribuição não-degenerada G, quando $r \to +\infty$, então G é a função de distribuição de valores extremos G_γ dada por*

$$G_\gamma(x) = exp(-\left(1+\gamma x\right)^{-\frac{1}{\gamma}}) \quad , x \in \mathbb{R},$$

para algum γ real, coincidindo com a denominada distribuição de Gumbel quando $\gamma = 0$, com a distribuição de Fréchet quando $\gamma > 0$ e com a distribuição de Weibull quando $\gamma < 0$.

No caso concreto da estatística T_r, o Teorema 2.6.1 poderia contribuir para estabelecer uma distribuição aproximação de T_r quando o número de categorias r fosse suficientemente elevado.

Com base no resultado seguinte, conseguiremos rapidamente deduzir que a função de distribuição de cada uma das componente de χ_i de T_r, qui-quadrado com um grau de liberdade, pertence ao domínio de atracção da Gumbel.

Teorema 2.6.2 *Se F é uma função de distribuição contínua com limite superior de suporte infinito, e se*

$$\lim_{x\to+\infty}\Big(\frac{1-F}{F'}\Big)'(x)=\gamma\,,$$

então G_γ é a função de distribuição limite de $F^r(a_rx+b_r)$, quando $r\to+\infty$, para convenientes constantes de normalização $a_r>0$ e $b_r\in\mathbb{R}$.

Na realidade, sendo a função densidade de probabilidade da distribuição de qui-quadrado com um grau de liberdade dada por:

$$f(x)=\frac{e^{-\frac{x}{2}}x^{-\frac{1}{2}}}{\sqrt{2\pi}}\;,\;\;x>0\,,$$

vem:

$$\begin{aligned}
\lim_{x\to+\infty}\Big(\frac{1-F}{F'}\Big)'(x) &= \lim_{x\to+\infty}1+\frac{f'(1-F)}{f^2}(x)\\
&= \lim_{x\to+\infty}1+\frac{f(-1-x^{-1})(1-F)}{f^2}(x)\\
&= \lim_{x\to+\infty}1-(1+x^{-1})\frac{1-F}{f}(x),\\
&= 1-\lim_{x\to+\infty}\frac{1-F(x)}{f(x)},\\
&= 1-\lim_{x\to+\infty}\frac{-f(x)}{f(x)(-1-x^{-1})}=1-1=0,
\end{aligned}$$

satisfazendo a condição do Teorema 2.6.2 com $\gamma=0$, o que prova ser a distribuição de Gumbel a distribuição limite do máximo normalizado de n variáveis independentes com distribuição de qui-quadrado com um grau de liberdade.

Consequentemente, existem sucessões de constantes de normalização, $a_r > 0$ e b_r reais, tais que

$$\begin{aligned} & \lim_{r\to+\infty}\lim_{n\to+\infty} P\big(\max(\chi_1^2,\chi_2^2,\ldots,\chi_{r-1}^2) \le a_{r-1}t + b_{r-1}\big) \\ = & \lim_{r\to+\infty} F^{r-1}(a_{r-1}t + b_{r-1}) \\ = & \exp(-\exp(-t)), \forall t \in \mathbb{R} \end{aligned} \tag{2.10}$$

onde, segundo [13], as constantes de normalização podem ser dadas por b_r tal que $1 - F(b_r) = \dfrac{1}{r}$ e $a_r = \dfrac{1}{rF'(b_r)}$.

Na prática, tendo em conta (2.9), quando n é elevado rejeitamos H_0, a um nível de significância α, se o valor observado de $T_r = max(\chi_1^2, \chi_2^2, \ldots, \chi_{r-1}^2)$ é maior que o quantil t_r, com $F^{r-1}(t_r) = 1 - \alpha$.
Se o tamanho da amostra n e o número de categorias r são ambos elevados, então podemos tomar o comportamento limite,

$$\lim_{r\to+\infty}\lim_{n\to+\infty} P(T_r \le \underbrace{a_{r-1}t + b_{r-1}}_{t_r}) = \exp(-\exp(-t)),$$

de forma a decidir sobre a rejeição de H_0. Tomando $t_r = a_{r-1}t + b_{r-1}$, isto é, $t = \dfrac{t_r - b_{r-1}}{a_{r-1}}$, a estatística T_r conduz à rejeitar H_0, a um nível de significância α, se $T_r > t_r$, onde t_r é determinado de modo que:

$$\begin{aligned} \alpha &= P(\text{rejeitar}H_0|\ H_0\text{verdadeira}) \\ &= P(T_r > t_r) \\ &= 1 - P(T_r \le t_r) \\ &= 1 - \exp\Big(-\exp\Big(-\frac{t_r - b_{r-1}}{a_{r-1}}\Big)\Big). \end{aligned}$$

Aplicando duas vezes o logaritmo, vem

$$-\exp\Big(-\frac{t_r - b_{r-1}}{a_{r-1}}\Big) = \ln(1-\alpha)$$

e portanto,

$$-\frac{t_r - b_{r-1}}{a_{r-1}} = \ln(-\ln(1-\alpha))$$

donde,

$$t_r = -\ln(-\ln(1-\alpha))a_{r-1} + b_{r-1}.$$

Assim, o teste conduz a rejeitar H_0 se $T_r > a_{r-1}(-\ln - \ln(1-\alpha)) + b_{r-1}$. Nesse caso, podemos identificar quais as categorias A_i que têm uma componente χ_i^2 maior que t_r, e assim, serem responsáveis por uma eventual rejeição da hipótese de homogeneidade (2.1).

No entanto, como observaremos no Capitulo 4, o valor desta estatística de teste T_r depende da ordenação das r categorias.

Capítulo 3

Testes de Homogeneidade em Amostras Emparelhadas

3.1 Introdução

O teste de McNemar é um procedimento estatístico, não paramétrico, usado em dados nominais e construído para testar a homogeneidade entre duas marginais em tabelas $k \times k$.

Este teste, introduzido por Quinn McNemar em 1947, foi inicialmente aplicado a tabelas de contingência 2×2 tendo como objectivo comparar a distribuição de duas variáveis emparelhadas, isto é, em que os dados são pares de observações correspondentes a um vector aleatório bi-dimensional e onde cada componente do vector tem exactamente dois resultados possíveis.

No caso dos resultados possíveis daquele vector aleatório bi-dimensional poderem ser classificados em múltiplas categorias, recorre-se ao teste estatístico proposto por Stuart(1955) e Maxwell(1970) que, sendo uma generalização do teste de McNemar, é conhecido por teste de McNemar generalizado ou de Stuart-Maxwell.

Considerando A e B duas variáveis nominais emparelhadas de uma determinada população, em que os resultados possíveis de cada uma das componentes que constituem os pares de dados estão distribuídos pelas mesmas r categorias, a tabela de contingência que resulta da classificação de n observações ou indivíduos nas $r \times r = r^2$ categorias

cruzadas tem a forma da Tabela 2.1 com $c = r$.

Neste caso, as variáveis emparelhadas A e B seguem, cada, uma distribuição multinomial, onde p_{ij} é a probabilidade conjunta do par (A, B) ocupar a linha i e a coluna j, para $i, j = 1, \ldots, r$, e $p_{i.}$ e $p_{.j}$ as distribuições marginais (totais marginais) das linhas e das colunas, respectivamente. Assim,

$$p_{i.} = \sum_{j=1}^{r} p_{ij}, \quad p_{.j} = \sum_{i=1}^{r} p_{ij} \quad \text{e} \quad \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{r} p_{.j} = \sum_{i=1}^{r} p_{ij} = 1.$$

Por definição, existe homogeneidade marginal entre duas variáveis emparelhadas A e B se se verificar a igualdade entre as proporções marginais das linhas e as proporções das colunas correspondentes. Estamos assim, perante o problema genérico de testar a hipótese nula:

$$H_0 : p_{k.} \quad = \quad p_{.\,k} \quad , \quad k = 1, \ldots, r \quad , \tag{3.1}$$

contra a hipótese alternativa:

$$H_1 : \exists_k : p_{k.} \quad \neq \quad p_{.k}.$$

Denotando $d_k = p_{k.} - p_{.\,k}$ e $\mathbf{d} = (d_1, \ldots, d_{r-1})$, é redundante incluir d_r no vector $\mathbf{d}$, uma vez que temos $\sum_{i=1}^{r} p_{i\,.} = 1$ (ou $\sum_{j=1}^{r} p_{.\,j} = 1$) e $\sum_{k=1}^{r} d_k = 0$. Por conseguinte, testar a hipótese (3.1) é equivalente a testar:

$$H_0 : p_{k.} \quad = \quad p_{.\,k} \quad , \quad k = 1, \ldots, r-1. \tag{3.2}$$

3.1.1 Teste de McNemar em tabelas 2×2

Quando $r = 2$, estamos perante a hipótese:

$$H_0 : p_{1.} = p_{.\,1} \tag{3.3}$$

sendo a estatística de teste de McNemar, neste caso, dada por:

$$Z = \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}}, \tag{3.4}$$

a qual segue assimptóticamente uma distribuição de qui-quadrado com um grau de liberdade [2], e sobre a qual em amostras de pequena dimensão deve ser aplicada a correcção à continuidade de Yates. Em [4] é referido que quando $(n_{21} + n_{12})$ é maior que dez, aquela aproximação é satisfatória.

Rejeitamos a hipótese (3.3), ao nível de significância α, se o valor observado da estatística de McNemar, Z_{obs}, for maior que o quantil de ordem $(1 - \alpha)$ de uma distribuição de qui-quadrado com um grau de liberdade, isto é,

$$Z_{obs} > \chi^2_{1;\,(1-\alpha)}.$$

Uma observação interessante quando se faz a interpretação da estatística (3.4), diz respeito ao facto dos elementos da diagonal principal da tabela de contingência 2×2 associada não contribuirem com qualquer informação para o valor daquela estatística.

3.1.2 Generalização do Teste de McNemar

Para testar a hipótese (3.2) temos de analisar o vector média e a matriz de covariâncias do vector $\mathbf{d}$ [14].

Sob a hipótese de homogeneidade marginal, tem-se que E($\mathbf{d}$)=0, e a matriz de covariâncias do vector $\sqrt{n}\mathbf{d}$, $\mathbf{V}$, tem dimensão $(r-1)(r-1)$, em que n é a dimensão total da amostra.

Stuart(1955) propôs a estatística Z_0 dada pela seguinte expressão

$$Z_0 = n\widehat{\mathbf{d}}^t\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{d}}\ ,$$

onde $\widehat{\mathbf{d}}$ representa um estimador de $\mathbf{d}$, com valores $(\widehat{d}_1, \ldots, \widehat{d}_{r-1})$ dados por $\widehat{d}_k = \sum_{i \neq k} n_{ki} - \sum_{i \neq k} n_{ik}$, e $\widehat{\mathbf{V}}$ um estimador da matriz de covariâncias do vector $\sqrt{n}\widehat{\mathbf{d}}$, com valores $\widehat{\mathbf{V}} = [\widehat{v}_{st}]_{s,t=1\ldots,r-1}$, dados por

$$\widehat{v}_{st} = -\frac{(n_{st} + n_{ts})}{n}, \text{ para } t \neq s \text{ e } t, s = 1, \ldots, r-1,$$

$$\widehat{v}_{ss} = \frac{n_{s.} + n_{.s} - 2n_{ss}}{n}, \quad \text{para} \quad s = 1, \ldots, r-1.$$

A estatística Z_0 do teste generalizado de McNemar, ou teste de Stuart-Maxwell, segue assimptoticamente uma distribuição de qui-quadrado com $r-1$ graus de liberdade [4]. Assim, rejeitamos a hipótese (3.2), ao nível de significância α, se o valor observado da estatística de Stuart, $Z_{0,obs}$, for maior que o quantil de ordem $(1-\alpha)$ de uma distribuição de qui-quadrado com $r-1$ graus de liberdade, isto é,

$$Z_{0,obs} > \chi^2_{(r-1);\,(1-\alpha)}.$$

Quando $r = 2$, a estatística do teste generalizado de McNemar, Z_0, é reduzida à estatística de McNemar dada por (3.4). No caso de $r = 3$, Walker (2002) desenvolveu uma fórmula que permite calcular o valor da estatística Z_0, e que é dada através da seguinte expressão

$$Z_0 = \frac{\overline{n}_{23}\widehat{d}_1^2 + \overline{n}_{13}\widehat{d}_2^2 + \overline{n}_{12}\widehat{d}_3^2}{2(\overline{n}_{12}\overline{n}_{23} + \overline{n}_{12}\overline{n}_{13} + \overline{n}_{13}\overline{n}_{23})}$$

onde $\overline{n}_{ij} = \dfrac{n_{ij} + n_{ji}}{2}$, para $i \neq j$.

3.1.3 Teste de Bhapkar

Bhapkar(1966), testou a hipótese de homogeneidade marginal (3.2) fazendo uso da normalidade assimptótica das proporções marginais, ficando este teste conhecido por teste de Bhapkar. O modo de construção desta estatística é semelhante ao da estatística generalizada de McNemar, residindo a maior diferença no cálculo dos elementos da matriz de covariância estimada, $\widehat{\mathbf{V}}$, sendo os seus elementos calculados da seguinte forma:

$$\widehat{v}_{st} = \frac{-(n_{st} + n_{ts}) - (n_{.s} - n_{s.})(n_{.t} - n_{t.})}{n}, \text{ para } t \neq s \text{ e } t, s = 1, \ldots, r-1,$$

$$\widehat{v}_{ss} = \frac{n_{s.} + n_{.s} - 2n_{ss} - (n_{.s} - n_{s.})^2}{n}, \text{ para } s = 1, \ldots, r-1.$$

A expressão da estatística de Bhapkar(1966) é igual à expressão de Z_0, isto é,

$$Z_1 = n\widehat{\mathbf{d}}^{\,t}\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{d}}\ ,$$

seguindo igualmente uma distribuição assimptótica de qui-quadrado com $r-1$ graus de liberdade, mas diferente de Z_0 no cálculo de $\widehat{\mathbf{V}}$.

Rejeitamos a hipótese (3.2), ao nível de significância α, se o valor observado da estatística Z_1, for maior que o quantil de ordem $(1-\alpha)$ de uma distribuição de qui-quadrado com $r-1$ graus de liberdade.

Ireland et al(1969) verificou existir a seguinte relação entre a estatística generalizada de McNemar e a estatística de Bhapkar:

$$Z_1 = \frac{Z_0}{1 - Z_0/n}.$$

Sendo os testes de Bhapkar e Stuart-Maxwell assimptóticamente equivalentes (Keefe, 1982), geralmente, o teste de Bhapkar (1966) é a alternativa mais potente ao teste de Stuart-Maxwell. Para um n grande, ambos produzem o mesmo valor de qui-quadrado. Uma vez que o teste de Bhapkar é mais potente, é o preferido.

Capítulo 4

Aplicações ao Caso em Estudo

4.1 Introdução

Neste capítulo são aplicados vários dos testes abordados no Capítulo 3, com o objectivo prático de analisar a existência de homogeneidade entre dois pares de espécies tidas biologicamente como muito semelhantes em termos de mapas de contextos. O que são mapas de contextos? Estes mapas são construídos no *Anaconda*[1]. Basicamente, para cada genoma sequenciado, o *Anaconda* importa o conjunto completo de ORFeoma de uma base de dados pública e converte-o numa tabela de contingência 64×64, relativa aos pares de codões justapostos existentes no ORFeoma. Com essa tabela, é testado a existência de não associação entre dois pares consecutivos usando a estatística de qui-quadrado de Pearson e é construída a matriz dos resíduos ajustados de Pearson [2]. O mapa de contexto de pares de codões corresponde a essa matriz dos resíduos ajustados de Pearson que, para uma melhor visualização, são convertidos numa escala de cores como ilustram as Figuras 4.2 e 4.4 onde são apresentados os mapas de contextos para as espécies *Homo sapiens*, *Pan troglodytes*, *Saccharomyces cerevisiae* e *Saccharomyces paradoxus* [3,12].

Há várias décadas que a comparação entre o comportamento dos seres humanos e dos chimpanzés é alvo de inúmeras investigações. Como o homem, o chimpanzé é um dos poucos animais que consegue reconhecer a própria imagem ao espelho, ou aprender

[1]Este software pode ser extraído no site http://bioinformatics.ua.pt/aplications/anaconda.

certos tipos de linguagens como a dos sinais.

Figura 4.1: Imagem referente às palmas das mãos e dos pés dos hominídeos e do homem.

Apesar do ramo evolutivo destas duas espécies se ter separado há cerca de quatro a sete milhões de anos atrás, os cientistas calculam que a diferença entre o ADN da espécie *Homo sapiens* (homem) e o ADN do *Pan troglodytes* (chimpanzé comum) é de apenas cerca de 2 %.

Na Figura 4.3[2] ilustra-se a espécie *Saccharomyces cerevisiae* que é uma levedura usada na produção de pão, cerveja, vinho, enzimas e produtos farmacêuticos. Por outro lado, a *Saccharomyces paradoxus* é uma levedura usada na produção de vinhos, estando intimamente relacionada com a *Saccharomyces cerevisiae*. Estes organismos unicelulares eucariontes[3], pertencentes ao grupo dos Fungos, são muito utilizados como modelo no estudo da Bioquímica, da Genética e da Biologia Celular na compreensão de processos celulares e moleculares de seres eucariontes. Apresentando como principais vantagens o facto de serem organismos unicelulares que, ao contrário dos organismos eucariontes complexos, são de fácil manutenção e manipulação em laboratório, e terem estruturas relativamente semelhantes à das células humanas (ambas eucariontes), tendo muitas das proteínas importantes na biologia humana sido estudadas através dos seus homólogos nas leveduras. Estima-se que o ramo evolutivo destas duas espécies se tenha separado à cerca de cinco a dez milhões de anos de atrás.

[2]Imagens retiradas da página da Internet com o seguinte endereço: www.bath.ac.uk/.../images/profiles/wheals2.gif.

[3]Seres eucariontes são organismos cujas células têm núcleo.

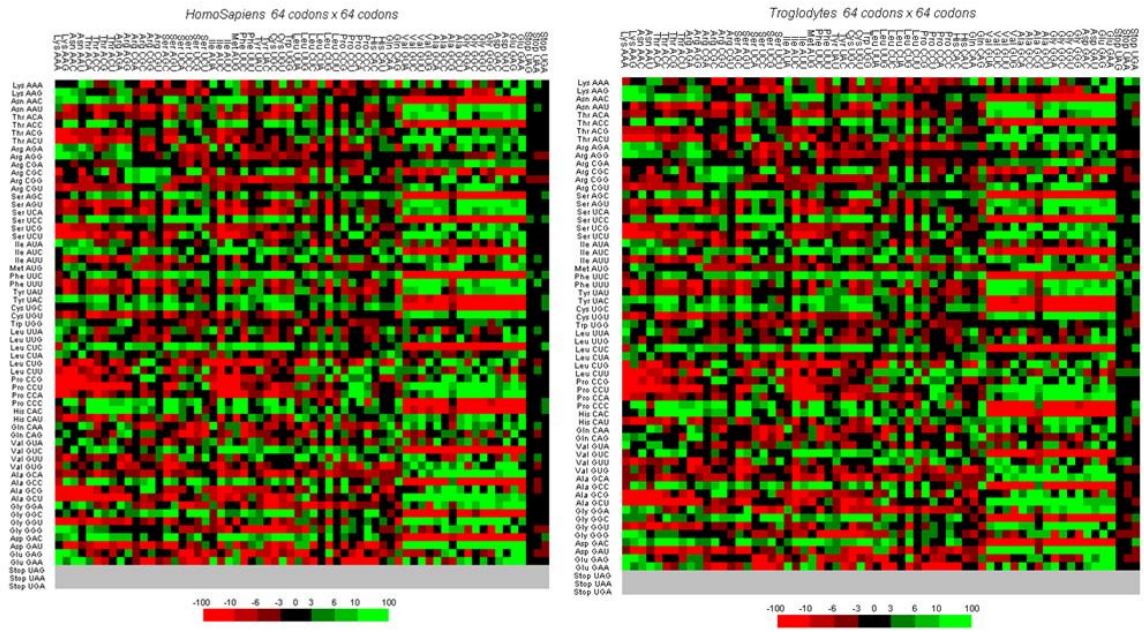


Figura 4.2: Mapas de contextos das espécies *Homo sapiens* e *Pan troglodytes*: Em termos matemáticos, cada mapa representa a matriz de resíduos ajustados de Pearson, resultante da identificação dos pares de codões responsáveis pela rejeição da hipótese de independência entre codões consecutivos nas sequências codificantes (ORFeoma) da espécies considerada. Os resíduos são transformados numa escala de cores de acordo com a escala anexa. O verde (vermelho) representa pares de codões preferidos (preteridos), no sentido em que se observam mais (menos) do que seria esperado se não existisse associação entre pares de codão consecutivos nas sequências codificantes.

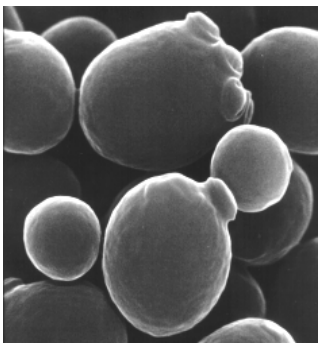


Figura 4.3: Imagens da *Saccharomyces cerevisiae*, vulgarmente denominada de levedura do pão.

4.2 Homogeneidade de Símbolos Genéticos

Aproveitando as potencialidades do software *Anaconda*, extraímos as sequências codificantes do genoma da cada uma daquelas quatro espécies de interesse no presente

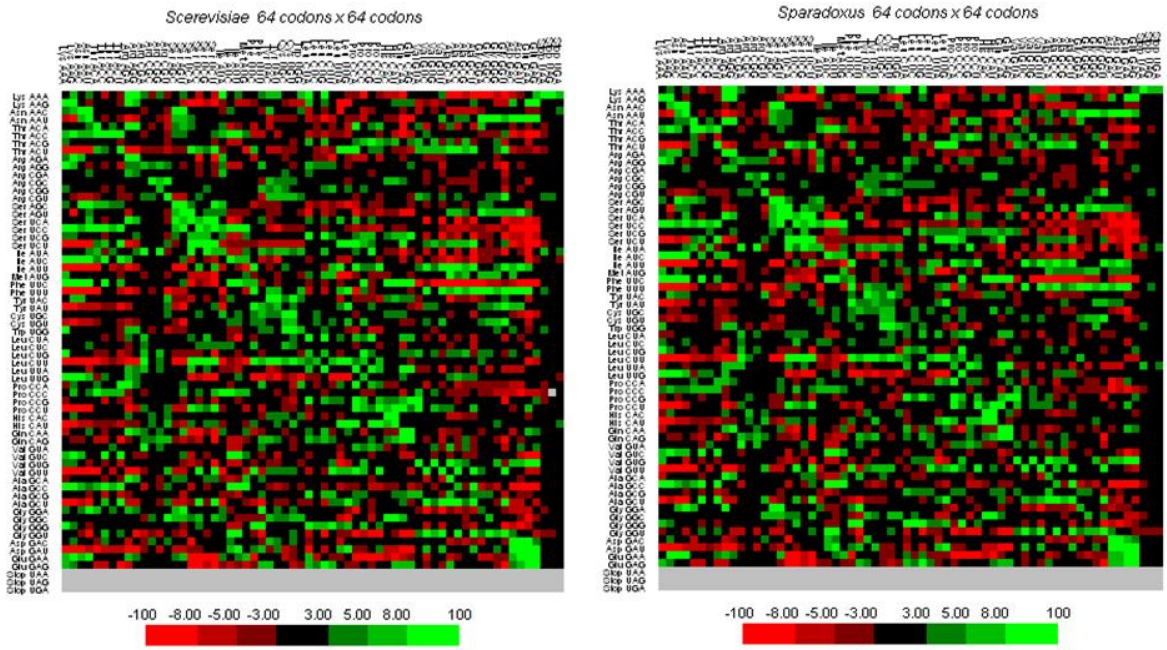


Figura 4.4: Mapas de contextos das espécies *S.cerivisae* e *S.paradoxus* - A grande semelhança observada entre estas imagens pode ser constatada, por exemplo, em relação à diagonal, em que ambas têm tendência para tons verde, podendo-se afirmar que os codões, na sua disposição sequencial, têm preferência pela justaposição com eles próprios.

trabalho no sítio http://www.ensembl.org/index.html, e construímos as tabelas de contingência (ver algumas em Anexo), relativas às contagens de vários contextos de símbolos genómicos nas sequências codificantes das quatro espécies, nomeadamente:

a) Pares de codões da forma $X_1X_2C - GX_5X_6$, onde X_1X_2C representa o primeiro codão do par com o terceiro nucleótido igual à Citusina, GX_5X_6 o segundo codão do par com o primeiro nucleótido igual à Guanina, e onde X_i pode representar qualquer nucleótido A, C, G ou U, na posição i do par.

b) Pares de codões da forma $X_1X_2X_3 - AX_5X_6$, onde $X_1X_2X_3$ representa o primeiro codão do par, AX_5X_6 o segundo codão do par com o primeiro nucleótido igual à Adenina, e onde X_i pode representar qualquer nucleótido A, C, G ou U, na posição i do par.

c) Pares de codões da forma $XYZ - XYZ$, isto é, pares de codões iguais, onde X, Y e Z pode representar qualquer nucleótido A, C, G ou U.

d) Todos os possíveis pares de codões, ou seja, pares da forma $X_1X_2X_3 - X_4X_5X_6$, onde X_i pode representar qualquer nucleótido A, C, G ou U, na posição i do par.

e) pares de codões da forma $X_1X_2U - AX_5X_6$, onde X_1X_2U representa o primeiro codão do par com o terceiro nucleótido igual ao Uracil, AX_5X_6 o segundo codão do par com o primeiro nucleótido igual à Adenina, e onde X_i pode representar qualquer nucleótido A, C, G ou U, na posição i do par.

Existindo quatro possíveis nucleótidos, as tabelas formadas nas situações **a)**, **b)**, **c)**, **d)** e **e)** têm, respectivamente, 256 ($= 4 \times 4 \times 1 \times 1 \times 4 \times 4$), 61 ($= 4 \times 4 \times 4$ - 3 codões de terminação), 61 ($= 4 \times 4 \times 4 \times 1 \times 1 \times 1$ - 3 codões de terminação), 3904 (($4 \times 4 \times 4$ - 3)$\times 4 \times 4 \times 4$) e 256 categorias em linha.
Testamos a igualdade de distribuição dos contextos **a)**, **b)** e **c)** entre o *Homo sapiens* e o *Pan troglodytes*, e dos contextos **d)**, **e)** e **c)** entre a *Saccharomyces cerevisiae* e a *Saccharomyces paradoxus*.

4.2.1 *Homo sapiens* versus *Pan troglodytes*

Relativamente as espécies *Homo sapiens* e *Pan troglodytes*, as situações **a)**, **b)** e **c)** correspondem à análise de uma tabela de contingência 256×2, 61×2 e 61×2, respectivamente.
Esses casos são escolhidos na sequência de estudos anteriores, [6], terem demonstrado *i)* haver uma associação negativa nos pares de codões da forma $X_1X_2C - GX_5X_6$ nas duas espécies, *ii)* existir uma maior predominância do nucleótico A (em qualquer posição do codão) em pares de codões justapostos onde existem diferenças significativas na distribuição nas duas espécies, e *iii)* haver uma predominância pela cor verde nas diagonais dos mapas de contextos.

Relativamente ao primeiro caso, os valores obtidos aplicando a estatística de Pearson ($\chi^2_{obs} = 803,84$) e da razão de verosimilhança ($G^2_{obs} = 796,64$) permitem concluir não existir homogeneidade entre as duas populações, em virtude de se tratarem de valores muito elevados (valores de $p-values$ quase nulos, $P(\chi^2 > 803,84); P(G^2 > 796,64) << 10^{-100}$). Conclui-se, não existir igualdade de proporções entre as espécies *Homo sapiens* e *Pan troglodytes* em relação aos pares de codões em que o primeiro codão termina com o nucleótido Citusina e em que o codão seguinte tem como nucleótido inicial a Guanina. No que diz respeito a aplicação da estatística da ϕ-divergência, o estimador mínimo foi calculado tendo em conta a medida da potência-divergência para $\lambda = -2, -\frac{1}{2}, 0, \frac{2}{3}, 1$ e 2, relativamente às funções $\phi_1(x) = \frac{1}{2}(x^{-1}-x)$, $\phi_1(x) = x\ln x - x + 1$ e $\phi_1(x) = \frac{1}{2}(x-1)^2$, tendo-se obtido os valores sumariados na seguinte tabela:

Estatística ϕ-Divergência						
	$\lambda = -2$	$\lambda = -\frac{1}{2}$	$\lambda = 0$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
$\phi_1(x) = \frac{1}{2}(x^{-1} - x)(\lambda = -2)$	783,85	785,13	786,21	788,26	789,57	794,73
$\phi_1(x) = x\ln x - x + 1(\lambda = 0)$	798,98	796,80	796,64	796,95	797,36	799,67
$\phi_1(x) = \frac{1}{2}(x-1)^2(\lambda = 1)$	808,84	804,70	803,84	803,20	803,12	803,92

Constatamos que os valores obtidos para as estatísticas são muito próximos dos obtidos aquando da aplicação das estatísticas de Pearson e da razão de verosimilhança, não se registando grande diferença no valor do λ utilizado para uma mesma função ϕ_1. No entanto, parece existir alguma relação entre o λ e a função $\phi_1(x)$ utilizada em relação aos resultados obtidos pela estatística. Por exemplo, em relação à função $\phi_1(x) = \frac{1}{2}(x^{-1} - x)(\lambda = -2)$, obtém-se o menor valor quando o estimador mínimo é calculado tomando $\lambda = -2$, o mesmo acontece em relação à função $\phi_1(x) = \frac{1}{2}(x-1)^2(\lambda = 1)$, em que o menor valor é obtido para $\lambda = 1$.

Todas aquelas estatísticas de teste revelam não existir homogeneidade entre as populações em estudo em relação à situação **a)**.

Tomemos agora a estatística de teste T_r, considerando as categorias em linha ordenadas segundo três maneiras distintas: de forma aleatória e ordenando as linhas pelas iniciais dos aminoácidos, correspondentes ao primeiro codão do par, de Z a A e de A a Z.

Na tabela que se segue, temos o valor da estatística de teste para as três diferentes ordenações das 256 categorias, assim como a identificação das categorias que apresentam um desvio significativo, ao nível de significância de 0,01, relativamente à hipótese de homogeneidade.

Valor da Estatística T_r		
$T_{256} = 33,41$	$T_{256} = 22,94$	$T_{256} = 18,94$
33,41 Gly GGC GCG	22,94 Val GUC GUC	18,94 Thr ACC GAG
32,82 Asp GAC GCC	20,93 Cys UGC GUC	18,75 Ser AGC GCG
27,17 Asn AAC GAG	17,54 Cys UGC GCC	17,60 Phe UUC GAC
25,90 Ser AGC GUG	17,14 Ser AGC GGG	16,73 Phe UUC GUA
25,29 His CAC GAA		
22,42 Pro CCC GAC		
20,82 His CAC GUU		
19,84 Ser UCC GGA		
19,16 Ser UCC GGC		
17,99 Tyr UAC GAC		
16,94 Leu CUC GGA		

Rapidamente se constata a dependência da estatística T_r na ordenação das categorias na tabela. Porém, para qualquer uma das três ordenações concluí-se pela não existência de homogeneidade, em virtude dos valores obtidos para a estatística T_r seres superiores ao valor do quantil t_r (=16,68, com $r = 256$), de ordem 0,99, destacando-se uma maior presença dos codões GCG, GCC, GAG, GAC e AGC nos pares responsáveis pela rejeição da hipótese de homogeneidade.

Em relação ao segundo caso - situação **b)** - os resultados obtidos aplicando as estatísticas de teste de Pearson, da razão e da ϕ-divergência, conduzem a valores elevados dessas estatísticas, como se pode observar nas duas tabelas que se seguem, pelo que as conclusões anteriormente retiradas em relação ao primeiro caso, de não homogeneidade, também se mantêm neste segundo caso.

Estatística de Pearson	Estatística da Razão de Verosimilhança
2011,56	2008,97

Estatística ϕ-Divergência						
	$\lambda=-2$	$\lambda=-\frac{1}{2}$	$\lambda=0$	$\lambda=\frac{2}{3}$	$\lambda=1$	$\lambda=2$
$\phi_1(x)=\frac{1}{2}(x^{-1}-x)(\lambda=-2)$	2004,76	2005,64	2006,32	2007,55	2008,29	2011,09
$\phi_1(x)=x\ln x-x+1(\lambda=0)$	2010,52	2009,06	2008,97	2009,14	2009,36	2010,57
$\phi_1(x)=\frac{1}{2}(x-1)^2(\lambda=1)$	2014,69	2012,05	2011,56	2011,21	2011,16	2011,57

Consideramos várias ordenações aleatórias das 61 categorias. Verificamos que para todas elas o valor obtido aplicando a estatística de teste, T_r, conduz sempre à não existência de homogeneidade, dado o valor observado para a estatística ser superior ao do quantil t_r (=13,81, $r=61$) de ordem 0,99. Dado também se ter registado sempre um elevado número de categorias (mais ou menos metade das categorias) com um valor da componente superior a t_r, tal não nos possibilitou encontrar um padrão sobre potenciais responsáveis pela rejeição da hipótese de homogeneidade, à semelhança do que conseguimos na situação **a)** acima analisada.

Relativamente ao estudo da homogeneidade da distribuição dos pares de codões iguais, situação **c)**, tanto aplicando a estatística de Pearson, a da razão de verosimilhança, como as estatísticas da ϕ-divergência, todas conduzem à decisão de rejeitar a hipótese de homogeneidade, ao nível de significância de 0,01. Os valores observados para essas estatísticas encontram-se sumariados nas seguintes tabelas.

Estatística de Pearson	Estatística da Razão de Verosimilhança
740,91	741,28

Estatística ϕ-Divergência						
	$\lambda=-2$	$\lambda=-\frac{1}{2}$	$\lambda=0$	$\lambda=\frac{2}{3}$	$\lambda=1$	$\lambda=2$
$\phi_1(x)=\frac{1}{2}(x^{-1}-x)(\lambda=-2)$	743,51	744,62	745,47	746,97	747,86	751,14
$\phi_1(x)=x\ln x-x+1(\lambda=0)$	743,24	741,40	741,28	741,49	741,75	743,14
$\phi_1(x)=\frac{1}{2}(x-1)^2(\lambda=1)$	744,79	741,50	740,91	740,49	740,44	740,90

A aplicação da estatística de teste T_r vem reforçar a rejeição da hipótese de homogenidade. Para várias ordenações consideradas das 61 categorias, o valor obtido foi sempre superior ao do quantil t_r (13,81) de ordem 0,99. Na seguinte tabela temos os valores observados de T_r assim como a identificação dos codões (em pares de codões iguais) responsáveis pelos valores observados para T_r para três ordenações consideradas: a primeira e a terceira ordenação das 61 categorias obtidas de forma aleatória, e a segunda ordenando as iniciais dos aminoácidos de Z a A.

Valor da Estatística T_r		
$T_{64} = 156,33$	$T_{64} = 99,64$	$T_{64} = 94,98$
156,33 Glu GAG	99,64 Glu GAG	94,98 Pro CCA
60,98 Lys AAG	89,04 Ala GCA	73,57 Thr ACU
58,58 Ala GCU	63,57 Ser AGC	63,41 Ile AUC
42,86 Leu CUG	44,11 Ala GCG	40,64 Val GUU
37,40 Gln CAG	39,45 Val GUG	35,15 Arg CGA
36,64 Val GUG	38,63 Leu UUG	34,88 Phe UUU
35,01 Ala GCA	32,99 Gln CAA	33,35 Trp UGG
30,58 Arg CGG	30,77 Ser UCA	32,63 Tyr UAC
25,85 Ser UCU	30,29 His CAC	31,27 Met AUG
23,79 Glu GAA	27,42 Leu CUG	29,72 Arg CGG
23,53 Gln CAA	23,21 Pro CCU	27,02 His CAU
23,32 Pro CCC	20,36 His CAU	26,02 Asn AAC
20,20 Gly GGA	20,36 His CAU	26,02 Tyr UAU
14,17 Arg AGA	18,63 Thr ACC	19,41 Thr ACA
14,04 Arg AGG		19,12 Gln CAG
		15,87 Glu GAA

Em relação às categorias que apresentam um desvio relativamente à hipótese de homogeneidade, salienta-se o facto dos codões GAG, CUG, CAG, CAU, GUG, GCA, CGG, GAA e CAA aparecerem mais frequentemente nos pares responsáveis pela rejeição em duas das três aplicações. Outra curiosidade, é o facto do núcleótido A ser o que está

presente em maior número nos codões "desviantes"e o nucleótido U o que aparece menos.

4.2.2 *S. cerevisiae* versus S. *paradoxus*

Em relação a estas duas espécies, analisamos tabelas de contingência 3904×2, 256×2 e 61×2 relativas, respectivamente, às situações de contextos **d)**, **e)** e **c)** atrás apontadas. Relativamente à situação **d)**, as estatísticas de Pearson, da razão de verosimilhança e T_r, esta última para três diferentes ordenações das 3904 categorias (a primeira ordenação obtida aleatoriamente, a segunda ordenando as iniciais dos aminoácidos de A a Z e a terceira de Z a A), produzem os seguintes resultados:

Estatística de Pearson	Estatística da Razão de Verosimilhança	Estatística T_r
3612,15	3613,47	15,02 16,17 14,39

Da aplicação destas três estatísticas, observa-se a existência de homogeneidade entre estas duas espécies, com um nível de significância de 0,01, em virtude de $\chi^2_{3903\,:\,0,99} = 4111,47$ e $t_{3904} = 21,99$.

A aplicação da estatística da ϕ-divergência vem confirmar a decisão no sentido da homogeneidade.

Estatística ϕ-Divergência						
	$\lambda = -2$	$\lambda = -\frac{1}{2}$	$\lambda = 0$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
$\phi_1(x) = \frac{1}{2}(x^{-1} - x)(\lambda = -2)$	3607,66	3614,92	3622,21	3636,87	3645,66	3675,88
$\phi_1(x) = x\ln x - x + 1(\lambda = 0)$	3629,65	3614,61	3613,47	3612,74	3614,98	3626,29
$\phi_1(x) = \frac{1}{2}(x-1)^2(\lambda = 1)$	3650,35	3622,10	3612,15	3607,76	3607,28	3610,95

Podemos assim concluir, ao nível de significância de 1%, existir uma igualdade entre as proporções de cada par de codões justapostos que se encontram nas sequências codificantes da *Saccharomyces cerevisiae* e da *Saccharomyces paradoxus*.

Assim sendo, esperamos que os resultados para a situação **e)** acima descrita , e apresentados nas duas tabelas que se seguem, conduzam à não rejeição de homogeneidade, o que efectivamente se observa, ao nível de significância de 0,01, em virtude

de $\chi^2_{255\,:\,0,99} = 310,46$ e $t_{256} = 16,68$. Na aplicação da estatística T_r, três diferentes ordenações obtidas de forma aleatória são consideradas.

Estatística de Pearson	Estatística da Razão de Verosimilhança	Estatística T_r
205,66	205,64	8,40 6,96 8,63

Estatística ϕ-Divergência						
	$\lambda = -2$	$\lambda = -\frac{1}{2}$	$\lambda = 0$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
$\phi_1(x) = \frac{1}{2}(x^{-1} - x)(\lambda = -2)$	205,45	205,66	205,83	206,13	206,31	206,97
$\phi_1(x) = x\ln x - x + 1(\lambda = 0)$	206,02	205,66	205,64	205,68	205,73	206,01
$\phi_1(x) = \frac{1}{2}(x-1)^2(\lambda = 1)$	206,43	205,78	205,66	205,58	205,57	205,66

Em relação à análise da diagonal dos mapas de contexto das duas espécies (situação **c)**), efectivamente as quatro estatísticas conduzem à existência de homogeneidade entre pares de codões iguais, ao nível de significância de 0,01 ($\chi^2_{60\,:\,0,99} = 88,38$ e $t_{61} = 13,81$), tendo as estatísticas apresentado os seguintes valores:

Estatística de Pearson	Estatística da Razão de Verosimilhança	Estatística T_r
82,98	82,94	7,35 6,54 14,32

Estatística ϕ-Divergência						
	$\lambda = -2$	$\lambda = -\frac{1}{2}$	$\lambda = 0$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
$\phi_1(x) = \frac{1}{2}(x^{-1} - x)(\lambda = -2)$	82,70	82,92	83,10	83,41	83,60	84,29
$\phi_1(x) = x\ln x - x + 1(\lambda = 0)$	83,34	82,97	82,94	82,99	83,04	83,33
$\phi_1(x) = \frac{1}{2}(x-1)^2(\lambda = 1)$	83,77	83,10	82,98	82,89	82,88	82,97

Refira-se que, apenas se verifica a rejeição da hipótese de homogeneidade aquando da aplicação da estatística T_r ordenando as 61 categorias pelas iniciais dos aminoácidos de A a Z, e na qual se destaca o codão GCC como o único responsável pela rejeição da homogeneidade. Nas restantes ordenações, obtidas de forma aleatória, verifica-se sempre a homogeneidade.

4.3 Homogeneidade de Preferência

Para além da análise comparativa de distribuições em termos de alguns contextos acima efectuada, analisamos agora a distribuição dos contextos em termos de cor nos mapas, ou seja, em termos de preferência e preterência de pares de codões consecutivos.
Efectuando a contagem de todos os pares de codões preferidos (resíduo ajustado de Pearson > 3, cor verde no mapa de contextos), preteridos (resíduo ajustado de Pearson $<$-3, cor vermelha no mapa de contextos) e indiferentes (resíduo ajustado de Pearson entre 3 e -3, cor preta no mapa de contextos) nas sequências codificantes das espécies *Saccharomyces cerevisiae*, *Homo sapiens* e *Pan troglodytes*, construímos a seguinte tabela de contingência 3×3:

Cor de Contexto	S.cerevisiae	Homo sapiens	Pan troglodytes	Total Marginal
Vermelho	694202	3296968	2241468	6232638
Preto	927413	2608762	1171053	4707228
Verde	923128	6747168	4649416	12319712
Total Marginal	2544743	12652898	8061937	23259578

Com base na Secção 2.3.3, podemos obter uma partição da estatística de Pearson em tantas componentes independentes quanto o número de graus de liberdade da estatística correspondendo, cada uma destas componentes, a uma tabela 2×2 tirada da tabela inicial, e à qual está associado um qui-quadrado com um grau de liberdade.
As quatro tabelas 2×2 que definem uma partição do χ^2 da tabela de contingência 3×3 são:

Tabela 1			**Tabela 2**	
694202	3296968		3991170	2241468
927413	2608762		3536175	1171053
Tabela 3			**Tabela 4**	
1621615	5905730		7527345	3412521
923128	6747168		7670296	4649416

Calculando o valor da estatística de Pearson sobre a tabela de contingência original 3×3, e das estatísticas (2.5) sobre cada uma das quatro tabelas 2×2, respectivamente, obtemos os seguintes valores:

Tabelas				
Original	**1**	**2**	**3**	**4**
634832,68	132461,99	145523,76	247204,69	109642,24

De onde se conclui, não existir homogeneidade de preferência na distribuição dos contextos em termos de cor, entre pares de codões consecutivos das três espécies, ao nível de significância 0,01, em virtude de $\chi^2_{4\,:\,0,99} = 13,28$ e $\chi^2_{1\,:\,0,99} = 6,63$.

Esta partição da estatística de Pearson permite-nos também observar, que o valor mais significativo é o da terceira componente, à qual está associada a terceira tabela 2×2. No entanto, torna-se difícil interpretar os resultados obtidos, e identificar quais os principais responsáveis pela rejeição da homogeneidade, pois a primeira categoria desta terceira tabela resulta da junção das duas primeiras categorias da tabela original em relação às espécies, a *Saccharomyces cerevisiae* e o *Homo sapiens*.

Por último, mostramos que a construção das tabelas 2×2, que definem a partição da estatística do qui-quadrado, dependem da ordem pela qual as categorias estão organizadas na tabela original.

Na realidade, alteremos a ordem das categorias da tabela 3×3 inicial produzindo a seguinte tabela equivalente:

Cor de Contexto	S.cerevisiae	Homo sapiens	Pan troglodytes	Total Marginal
Verde	923128	4649416	6747168	12319712
Vermelho	694202	2241468	3296968	6232638
Preto	927413	1171053	2608762	4707228
Total Marginal	2544743	8061937	12652898	23259578

As quatro tabelas 2×2 que definem a partição do χ^2 da tabela de contingência 3×3 são agora:

Tabela 1			**Tabela 2**	
923128	4649416		5572544	6747168
694202	2241468		2935670	3296968
Tabela 3			**Tabela 4**	
1617330	6890884		8508214	10044136
927413	1171053		2098466	2608762

Aplicando a estatística de Pearson à tabela de contingência 3×3, e as estatísticas (2.5) a cada uma das quatro tabelas 2×2, respectivamente, obtivemos os seguintes valores:

Tabelas				
3×3	**1**	**2**	**3**	**4**
634832,68	50860,22	5826,96	575662,13	2483,36

Constatando-se assim, que o valor das componentes depende da ordem pela qual as categorias se encontram na tabela de contingência inicial. Porém, nos dois casos a soma dessas quatro componentes de qui-quadrado coincide com o valor da estatística de Pearson sobre a tabela 3×3 inicial:

$$\begin{aligned} 634832,68 &= 132461,99 + 145523,76 + 247204,69 + 109642,24 \\ &= 50860,22 + 5826,96 + 575662,13 + 2483,36 \end{aligned}$$

Nas duas situações, as quatro componentes definem uma partição da estatística de Pearson.

Capítulo 5

Conclusão

Apesar das espécies *Homo sapiens* e *Pan troglodytes* serem consideradas biologicamente muito semelhantes em termos de mapas de contextos, a aplicação das quatro estatísticas de teste abordadas no Capítulo 2 levam à rejeição da existência de homogeneidade, entre este par de espécies, relativamente aos contextos considerados, nomedamente, quando o codão fixo tem como último nucleótido a Citusina e o nucleótido inicial do codão seguinte a Guanina, quando o nucleótido inicial do codão seguinte é a Adenina, e relativamente aos codões que ocupam a diagonal dos respectivos mapas de contextos. Convém aqui ressalvar, que as tabelas de contingências foram construídas com base no sequenciamento conhecido daquelas duas espécies à data de Maio de 2008, estando apenas disponível cerca de um terço do ORFenoma do *Pan troglodytes*. Assim, as conclusões retiradas em relação a estas duas espécies podem ter sido influenciadas por tal facto.

Já em relação ao outro par de espécies, *Saccharomyces cerevisiae* e *Saccharomyces paradoxus*, a aplicação das quatro estatísticas vem confirmar a semelhança biológica tida como existente entre elas. Perante os contextos abordados, nomeadamente, em relação aos possíveis pares de codões justapostos que se podem encontrar nas sequências codificantes de cada espécie, quando o codão fixo tem como último nucleótido o Uracilo e o nucleótido inicial do codão seguinte a Adenina, e relativamente aos codões que ocupam a diagonal dos respectivos mapas de contexto, verifica-se sempre a existência de homogeneidade.

Na prática, as conclusões retiradas da aplicação das diferentes estatísticas revelam-se pouco diferentes.

Os valores observados para as estatísticas de Pearson e da razão de verosimilhança e para as estatísticas baseadas na medida de ϕ-divergência, vêm corroborar o que havia sido referido no Capitulo 2 relativo ao facto daquelas duas primeiras estatísticas se tratarem de casos particulares da estatística baseada na medida da ϕ-divergência.

Outro resultado que efectivamente se confirma na aplicação prática, é o facto do valor obtido pela estatística baseada no termo máximo, T_r, depender da ordenação das categorias na tabela de contingência. No entanto, aquando da rejeição da homogeneidade, esta estatística poderá ser útil na identificação de eventuais categorias com proporções significativamente não iguais nas populações sob comparação.

Quanto ao estudo feito em termos de preferência e preterência de pares de codões consecutivos nas sequências codificantes de três das espécies estudadas, a *Saccharomyces cerevisiae*, o *Homo sapiens* e o *Pan troglodytes*, conclui-se não existir homogeneidade na distribuição dos contextos em termos de cor, entre pares de codões consecutivos destas três espécies.

A não aplicação dos testes de homogeneidade em amostras emparelhadas nesta dissertação, fica-se a dever ao facto de inicialmente termos pensado em aplicá-los a dados de microarrays comparando a distribuição da amostra de genes (ser diferencialmente expresso e não diferencialmente expresso) não sujeitos e sujeitos a determinado choque térmico, não tendo sido porém possível obter esses dados para análise em tempo real.

Apêndice A

Teorema A.1: Seja $(N_{1j}, N_{2j}, \ldots, N_{r-1j})^t$ um vector aleatório de dimensão $r-1$ com distribuição multinomial de parâmetros $(n_{.j},\ (p_{1j}, \ldots, p_{r-1j})^t)$, referente à população j. Admitindo que é válida a hipótese de homogeneidade, a variável aleatória

$$\chi^2_{rj} = \sum_{i=1}^{r} \frac{\left(N_{ij} - \dfrac{n_{.j}N_{i.}}{n}\right)^2}{\dfrac{n_{.j}N_{i.}}{n}}$$

segue assimptoticamente (quando $n \to \infty$) a distribuição de qui-quadrado com $(r-1)$ graus de liberdade.

Prova A.1:

A prova que iremos apresentar trata-se da dedução obtida por Fisher e que se encontra em [11].

Para chegar à distribuição multinomial com parâmetros $(n_{.j},\ (p_{1j}, \ldots, p_{r-1j})^t)$, Fisher começa por considerar $(N_{1j}, N_{2j}, \ldots, N_{rj})$ variáveis aleatórias independentes com distribuição de Poisson de parâmetros $n_{.j}p_{ij}$, para $i = 1, \ldots, r$. Portanto, com função de probabilidade conjunta

$$P(n_{1j}, \ldots, n_{rj}) = \prod_{i=1}^{r} \frac{e^{-n_{.j}p_{ij}}(n_{.j}p_{ij})^{n_{ij}}}{n_{ij}!} = e^{-n_{.j}} n_{.j}^{n_{.j}} \prod_{i=1}^{r} \frac{p_{ij}^{n_{ij}}}{n_{ij}!}.$$

Sendo $\displaystyle\sum_{i=1}^{r} n_{.j}p_{ij} = n_{.j}$, pelo teorema 4.6 (ver Bento Murteira Vol I),

$$P\Big(\sum_{i=1}^{r} n_{ij} = n_{.j}\Big) = \frac{e^{-n_{.j}} n_{.j}^{n_{.j}}}{n_{.j}!}\,,$$

donde a distribuição condicionada,

$$\begin{aligned} P\Big(n_{1j}, \ldots, n_{rj} | \sum_{i=1}^{r} n_{ij} = n_{.j}\Big) &= \frac{P(n_{1j}, \ldots, n_{rj})}{P\Big(\displaystyle\sum_{i=1}^{r} n_{ij} = n_{.j}\Big)} \\ &= \frac{n_{.j}!}{n_{1j}! \ldots n_{rj}!} \prod_{i=1}^{r} p_{ij}^{n_{ij}} \end{aligned}$$

precisamente a multinomial que governa o comportamento probabilístico da estatística do qui-quadrado quando a hipótese é verdadeira.
Por outro lado, a estatística pode escrever-se em termos de quadrados de variáveis de Poisson estandardizadas

$$\chi^2_{rj} = \sum_{i=1}^{r} Z_{ij}^2 \quad , \quad Z_{ij} = \frac{n_{ij} - n_{.j}p_{ij}}{\sqrt{n_{.j}p_{ij}}} = \frac{n_{ij} - \frac{n_{.j}n_{i.}}{n}}{\sqrt{\frac{n_{.j}n_{i.}}{n}}} \quad , \quad i = 1, \ldots, r.$$

Logo, pelo teorema 5.26 (ver Bento Murteira Vol I)

$$Z_{ij} \sim N(0,1),$$

mostra que quando $n \to \infty$ e $n_{ij} \to \infty, i = 1, \ldots, r$, a estatística do qui-quadrado tende para a soma de quadrados de variáveis normais estandardizadas, isto é, pelo corolário 4.6 (ver Vol I de Bento Murteira) tem no limite uma distribuição qui-quadrado. Esta distribuição tem $(r-1)$ graus de liberdade e não r devido ao grau de liberdade perdido em consequência da restrição linear que existe entre as variáveis Z_{ij},

$$\sum_{i=1}^{r} (\sqrt{n_{.j}p_{ij}})\, Z_{ij} = 0$$

$$\Rightarrow (N_{1j} - n_{.j}p_{1j}) + (N_{2j} - n_{.j}p_{2j}) + \ldots + (N_{rj} - n_{.j}p_{rj}) = 0$$

$$\Rightarrow n_{.j} - n_{.j}(p_{1j} + \ldots + p_{rj}) = 0$$

$$\Rightarrow p_{rj} = -(p_{1j} + \ldots + p_{r-1j}).$$

□

Apêndice B

Espécies $X_1X_2X_3$ - AX_5X_6	*Homo sapiens*	*Pan troglogytes*	Total Marginal
Ala GCC	122343	42427	164770
Arg AGA	56976	23849	80825
Pro CCC	96718	33192	129910
Gly GGA	69334	27780	97114
Asp GAC	104383	36528	140911
Glu GAG	133976	47503	181479
Arg AGG	62734	24854	87588
Glu GAA	104598	40255	144853
Thr ACC	85647	30192	115839
Arg CGC	35493	11633	47126
Phe UUC	75968	26711	102679
Leu CUC	89000	31578	120578
Arg CGG	32158	10600	42758
Ala GCA	46258	18417	64675
Leu CUG	102508	36650	139158
Tyr UAC	56867	19910	76777
Ile AUC	79912	28460	108372
Gln CAA	48772	19109	67881
Val GUG	76332	27316	103648
Thr ACA	43620	17071	60691
Gly GGC	92514	33326	125840
Val GUC	68747	24582	93329
Pro CCA	47082	18230	65312
Ala GCU	31555	12429	43984
Ser UCA	33629	13175	46804
Cys UGC	54140	20716	74856
Lys AAA	83300	31480	114780
His CAU	25539	10091	35630
Ser UCU	27639	10867	38506
Asn AAC	81090	29427	110517
Val GUU	24040	9478	33518
His CAC	73559	27788	101347
Leu CUU	28708	11156	39864
Val GUA	23352	9167	32519
Ser UCC	78616	28635	107251
Ser AGU	25841	9983	35824
Gly GGU	21386	8277	29663
Trp UGG	52288	19714	72002
Leu UUG	37008	14055	51063
Lys AAG	116625	43512	160137
Pro CCG	14661	5090	19751
Pro CCU	29174	11132	40306
Ile AUU	31598	11982	43580
Phe UUU	35582	13455	49037
Ile AUA	28099	10177	38276
Arg CGA	17743	6348	24091
Cys UGU	22589	8572	31161
Thr ACU	24062	9112	33174
Leu UUA	26666	10065	36731
Leu CUA	26557	10018	36575
Asn AAU	39409	14767	54176
Tyr UAU	24094	8759	32853
Thr ACG	11912	4544	16456
Asp GAU	41472	15218	56690
Gln CAG	117169	43251	160420
Ser UCG	7790	2765	10555
Arg CGU	8002	2857	10859
Ala GCG	13271	4851	18122
Gly GGG	68808	25446	94254
Ser AGC	87875	32547	120422
Met AUG	63972	23681	87653
Total Marginal	3290760	1214760	4505520

Figura 1: Tabela de contingência em relação às espécies *Homo sapiens* e *Pan troglodytes* na análise do caso $X_1X_2X_3$ - AX_5X_6.

Espécies XYZ_XYZ	*Homo sapiens*	*Pan Troglodytes*	Total Marginal
Glu GAG	30773	9498	40271
Ala GCU	10659	4316	14975
Leu CUG	27131	8872	36003
Lys AAG	19302	7211	26513
Ala GCA	8910	3481	12391
Arg AGA	4814	2051	6865
Ala GCC	7755	2341	10096
Val GUG	12881	4144	17025
Gln CAG	20814	6970	27784
Arg CGG	3604	988	4592
Arg AGG	4879	1961	6840
Pro CCC	3753	1056	4809
Glu GAA	19113	6886	25999
Asp GAC	3734	1081	4815
Ser UCU	4489	1763	6252
Gly GGA	8014	2979	10993
Gln CAA	3201	1304	4505
Tyr UAC	3583	1055	4638
Leu CUU	3445	1386	4831
Asp GAU	9660	3519	13179
Ser AGC	12100	4343	16443
Ser UCA	3020	1183	4203
Met AUG	5237	1953	7190
Thr ACC	7999	2653	10652
Ile AUU	2746	1082	3828
Arg CGC	3547	1114	4661
Thr ACA	3378	1293	4671
His CAU	2321	919	3240
Gly GGC	6387	2126	8513
Pro CCG	3041	986	4027
Lys AAA	6623	2370	8993
Leu CUC	7030	2511	9541
Gly GGU	4273	1552	5825
Phe UUC	7480	2536	10016
Asn AAU	3139	1148	4287
His CAC	4160	1393	5553
Leu UUG	1846	693	2539
Pro CCU	7384	2517	9901
Ala GCG	3919	1313	5232
Pro CCA	6297	2147	8444
Leu CUA	774	311	1085
Cys UGU	1122	431	1553
Tyr UAU	1842	600	2442
Val GUA	944	366	1310
Ser UCC	7300	2504	9804
Val GUC	1344	437	1781
Asn AAC	6235	2139	8374
Gly GGG	2157	722	2879
Leu UUA	1127	419	1546
Ser AGU	1501	547	2048
Val GUU	2608	931	3539
Ile AUA	1118	365	1483
Arg CGU	413	124	537
Trp UGG	2895	988	3883
Ile AUC	8640	2993	11633
Thr ACU	1673	589	2262
Thr ACG	423	140	563
Arg CGA	420	140	560
Phe UUU	2335	806	3141
Cys UGC	5036	1753	6789
Ser UCG	487	169	656
Total Marginal	362835	126168	489003

Figura 2: Tabela de contingência em relação às espécies *Homo sapiens* e *Pan troglodytes* na análise do caso XYZ - XYZ.

Espécies XYZ - XYZ	S. paradoxus	S. cerevisae	Total Marginal
Ala GCA	729	895	1624
Ala GCC	458	485	943
Ala GCG	116	115	231
Ala GCU	1730	2058	3788
Arg AGA	1848	2158	4006
Arg AGG	287	299	586
Arg CGA	19	21	40
Arg CGC	40	50	90
Arg CGG	18	23	41
Arg CGU	230	257	487
Asn AAC	2047	2398	4445
Asn AAU	3635	4255	7890
Asp GAC	1190	1349	2539
Asp GAU	4665	5444	10109
Cys UGC	104	122	226
Cys UGU	222	252	474
Gln CAA	2592	2864	5456
Gln CAG	1110	1224	2334
Glu GAA	6394	7499	13893
Glu GAG	1399	1605	3004
Gly GGA	359	402	761
Gly GGC	326	340	666
Gly GGG	98	82	180
Gly GGU	2094	2466	4560
His CAC	238	261	499
His CAU	590	686	1276
Ile AUA	907	1080	1987
Ile AUC	953	1041	1994
Ile AUU	1808	2153	3961
Leu CUA	491	665	1156
Leu CUC	63	79	142
Leu CUG	404	406	810
Leu CUU	416	480	896
Leu UUA	1605	1865	3470
Leu UUG	1228	1396	2624
Lys AAA	4050	4850	8900
Lys AAG	3415	3762	7177
Met AUG	940	1046	1986
Phe UUC	911	1003	1914
Phe UUU	1313	1651	2964
Pro CCA	967	1133	2100
Pro CCC	99	86	185
Pro CCG	97	93	190
Pro CCU	539	677	1216
Ser AGC	455	501	956
Ser AGU	671	761	1432
Ser UCA	1178	1412	2590
Ser UCC	641	703	1344
Ser UCG	287	291	578
Ser UCU	1927	2327	4254
Thr ACA	882	1080	1962
Thr ACC	615	728	1343
Thr ACG	190	182	372
Thr ACU	1052	1364	2416
Trp UGG	329	389	718
Tyr UAC	660	767	1427
Tyr UAU	971	1122	2093
Val GUA	365	401	766
Val GUC	340	328	668
Val GUG	423	465	888
Val GUU	1403	1639	3042
Total Marginal	65133	75536	140669

Figura 3: Tabela de contingência em relação às espécies *Saccharomyces cerevisiae* e *Saccharomyces paradoxus* na análise do caso XYZ - XYZ.

Bibliografia

[1] Agresti, Alan, 2002. *Categorical Data Analysis*, 2nd Edition, Wiley.

[2] Agresti, Alan, 2007. *An Introduction to Categorical Data Analysis*, 2nd Edition, Wiley.

[3] Afreixo, Vera, 2002. Análise Estatística da Linguagem Genética, Departamento de Matemática da Universidade de Aveiro.

[4] Everitt, B.S., 1992. The Analysis of Contingency Tables, 2nd Edition, Chapman and Hall / CRC.

[5] Freitas, A. V., Pinheiro, M., Oliveira, J.L., Moura, G., Santos, M., 2005. A new limiting distribution for a statistical test for the homogeneity of two multinomial populations. *Proceedings of the Workshop in Statistic on Genonmics and Proteomics, CIM*, 27, 113-120.

[6] Freitas, A.,Duarte J., Pinheiro, M., Oliveira, J. L., Moura, G., Santos, M.,(2007). Homo sapiens versus Pan troglodytes: quão diferentes são?, in *Actas do XIV Congresso Nacional da Sociedade Portuguesa de Estatística*, Covilhã, Portugal.

[7] Gupta, A.K., Nguyen T., Pardo L., 2007. Residual analysis and outliers in log-linear models based on phi-divergence statistics. *Journal of Statistical Planning and Inference.*

[8] Kimball, A.W., 1954. Short-cut formulas for the exact partition of in contingency tables. *Biometrics*, 10, 452-458.

[9] Menéndez, M.L., Pardo, J.A., Pardo, L., Zografos, K., 2003. On tests of homogeneity based on minimum ϕ-divergence estimator with contraints. *Computational Statistics & Data Analysis*, 43, 215-234.

[10] Murteira, Bento, 1990. *Probabilidades e Estatística*, Vol. I, 2ª Edição, McGraw-Hill.

[11] Murteira, Bento, 1990. *Probabilidades e Estatística*, Vol. II, 2ª Edição, McGraw-Hill.

[12] Pinheiro, M., Afreixo, V., Moura, G., Freitas, A., Santos, M. A., Oliveira, J. L., 2006. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine*, 45 , 163-168.

[13] Reiss, R.-D and Thomas, M., 2007. Statistical Analysis of Extremes Values with Applications to Insurance, Finance, Hydrology and Other Fields, 3rd Edition, Birkhäuser.

[14] Xuezheng Sun, Zhao Yang, 2008. Generalized McNemar´s Test for Homogeneity of Marginal Distribuitions,*Proceedings of the SAS. GLOBAL FORUM 2008* Paper 382-2008.

[15] http://ferrari.dmat.fct.unl.pt/services/AnaliseDados/Contingencia.pdf

[16] http://74.125.77.132/search?q=cache:-DnQsN2ecQgJ:www.ine.pt/ngt_server /attachfileu.jsp